Sampling errors and design effects for poverty measures and other complex statistics

Vijay Verma, Gianni Betti

Working Paper n. 53, February 2005

Sampling errors and design effects for poverty measures and other complex statistics

Vijay Verma, Gianni Betti
University of Siena

ABSTRACT

The objective of the research reported here is to contribute to the development of methodology and practical tools for computing sampling errors and design effects for complex statistics based on complex sampling designs, specifically sampling error of measures of poverty and inequality. It is taken as given that for the 'typical' social surveys, based on reasonably large samples but with complex designs, the applicability of at least two broad approaches is generally well-established in the literature, namely the approaches based on Taylor linearisation, and on resampling such as Jackknife Repeated Replication (JRR). This research has concentrated on elaborating the necessary details and developing software for their practical application by researchers who are not necessarily experts in the field of complex variance estimation.

The linearisation approach to approximating variance of complex (non-linear) statistics is a long-established procedure. One of the important objectives of this working paper is to provide, in one place and derived in a uniform way, the linearised forms for a comprehensive set of income poverty and inequality measures. Where applicable, the JRR approach is by far the simpler technically. Apart from specifying the sample structure and defining appropriate 'computational units' for the purpose, the method merely involves repeated computation of the statistics (for which sampling errors are required) over sample replications. We develop and make available SAS routines for the efficient and accurate computation for complex measures of poverty and inequality based on survey data. Based on some empirical data, we compare the results from the linearisation and replication approaches, and explore the extent to which the results from the two procedures are similar.

Many statistics of interest can be written in the form of a ratio, but involving parameters which are themselves sample estimates – for instance the proportion poor, defined in terms of a poverty line itself estimated from the sample. What is the implication for the sampling error estimate of assuming these parameters to be constants, rather than sample dependent, i.e. of treating the complex statistic as an ordinary ratio? We compare the effects estimated under the two approaches.

Computing design effects (ratio of actual sampling error to that under equivalent simple random sampling, SRS) requires the additional step of estimating sampling errors under SRS. We propose and illustrate a practical procedure based on randomisation of units over structure of the sample and approximating the effect of sample weights on variance.

Our limited results suggest that for many measures of poverty and inequality, the JRR and linearisation methods give similar results. We generally recommend the former because of their analytical simplicity, better suited to the need of substantive researchers.

More importantly, we believe that the replication approach is readily extended to more complex designs and statistics (such as longitudinal measures in panel surveys), for which it is difficult to develop the linearised forms for variance estimation.

Contents

1. Introduction

This paper describes and illustrates procedures for the estimation of variances of complex statistics based on large sample surveys. The objective of this research is rather modest in theoretical terms. It is to *enhance the practice of routine computation of sampling errors* for complex statistics arising from large-scale samples of households and persons with complex designs. Specifically, we are concerned with sampling errors and design effects for the diversity of measures encountered in the analysis of income poverty and inequality based on sample data. The motivation for this work arises from the fact that information on sampling errors and design effects is very often not obtained or at least not reported or utilised in the analysis of substantive results of surveys. Various factors contribute to this situation, but we believe that (despite the fact that some general purpose software for the purpose has become available - see for instance, Brick and Morganstein, 1997), the lack of *easily accessible methods and tools for sampling error computations on a routine and large-scale basis* still remains a major contributing factor.

Given our specific and practical objectives, we will not review here the diversity of variance estimation approaches which have been developed for complex statistics from complex samples. Rather, we take it as given that for the 'typical' social surveys, based on reasonably large samples but with complex designs, the applicability of at least two broad approaches is generally well-established in the literature, namely the approaches based on (a) Taylor linearisation, and (b) on resampling such as Jackknife Repeated Replication (JRR). It is not our objective to explore the theoretical basis of these methods, but to concentrate on elaborating the necessary details and, to the extent possible, providing software for their practical application by researchers who are not necessarily experts in the field of complex variance estimation. We have developed these applications for most of the commonly used measures in the analysis of income poverty and inequality, at least in the conventional cross-sectional context. Work on extending the application to many other types of statistics, such as longitudinal indicators of poverty and deprivation, net changes and aggregation over time of such measures, indicators of economic activity and employment, etc, is in progress at the University of Siena.

The paper explores the following aspects.

(1) The linearisation approach to approximating variance of complex (non-linear) statistics is a long-established procedure. The basis of the approach is to use Taylor approximation to reduce non-linear statistics to a linear form, justified on the basis of asymptotic properties of large populations and samples. For each sample unit it seeks a *linearised 'indicative variable',* such that variance of the total of that variable approximates the variance of the complex statistic of interest. However, the derivations of the required linearised variables can be complex and, in our view, are often not available in the literature in a form which a practical researcher or statistician can readily adapt and apply. One of the important objectives of this paper is to provide, in one place and derived in a uniform way, the linearised forms for a comprehensive set of income poverty and inequality measures.

(2) In principle, the replication - in particular the JRR - method for variance estimation is straightforward. Where applicable, the approach is by far the simpler technically. Apart from specifying the sample structure and defining appropriate 'computational units' for the purpose, the method merely involves repeated computation of the estimate (for which sampling error is required) over different (often numerous) sample replications; variance of any statistic is estimated simply from variability in its estimates over the replications. The

form of the final variance estimation formula does not depend on the particular statistic involved. Apart from determining the appropriate procedure for constructing sample replications, details of the sampling design also do not complicate the picture.

The main shortcoming of the replication approach is the magnitude of the computational task involved in the repeated estimation of the statistic over a large number of full-scale replications of the sample - including some or all of the data adjustment and estimation steps (imputation, weighting, calibration, smoothing etc.) if their effect on variance is to be incorporated. This task is subject-matter specific. To this end we have undertaken the development of efficient and accurate SAS routines and macros for repeated computation of poverty and inequality measures, so as to encourage and facilitate routine computation of sampling errors for these statistics – something which is too often neglected by substantive researchers.

(3) Based on some empirical data, we compare the results from the linearisation and replication approaches. Are the results from the two procedures essentially the same for practical purposes, or are there significant or consistent differences in the sampling error estimates produced? There seems to be two schools among researchers, uniformly preferring the one or the other of the two approaches. We join this debate.

(4) Proceeding from estimates of sampling error to estimates of design effects (ratio of actual sampling error to that under equivalent simple random sampling, SRS) is essential for understanding the patterns of variation of sampling errors and the determinants of their magnitude, for smoothing and extrapolating the sampling error results for diverse statistics and population subclasses, and for evaluating the performance of the sampling design. Computing design effects requires the additional step of estimating sampling errors under simple random sampling. In practice this can be a far-from-trivial step for complex statistics for which explicit expressions are not available for the purpose. We propose and illustrate a practical procedure based on variance computed after randomisation of units over structure of the sample. This procedure permits approximate decomposition of the total design effect into the effect of sample weights, and that of clustering, stratification and other complexities of the sampling design.

(5) Many statistics of interest can be written in the form of a ratio, but involving parameters which are themselves sample estimates – for instance the proportion poor, defined in terms of a poverty line itself estimated from the sample. What is the implication for the sampling error estimate of assuming the estimates of these parameters to be constants, rather than sample dependent? We compare the effects estimated under the JRR and linearisation approaches. Necessary computational algorithms for the purpose are provided.

The main shortcomings of the JRR method are noted to be (i) the large scale of the computations which may be involved; and (ii) some doubts about applicability of the method to certain types of statistics, such as quantiles of the income distribution.

As to linearisation, the main limitations include (i) analytical complexity of the linearisation procedure; (ii) the impossibility, at least for the present, of obtaining the required linearisation forms for very complex statistics, such as coefficients in a logistic regression involving iterative estimation, or longitudinal measures of poverty; (iii) the methods may not be flexible enough to accommodate complexities in the design which may be involved in certain multiple, multiphase, or longitudinal samples; (iv) for certain poverty measures, an added complexity is that the linearised variance estimation formulas involve density functions at certain pints in the income distribution, which need to be estimated empirically.

2. Jackknife Repeated Replication (JRR) for variance estimation

The Jackknife Repeated Replication (JRR) is one of a class of methods for estimating sampling errors from comparisons among sample replications which are generated through repeated resampling of the same parent sample. Each replication needs to be a representative sample in itself and to reflect the full complexity of the parent sample. However, in so far as the replications are not independent, special procedures are required in constructing them so as to avoid bias in the resulting variance estimates. We prefer the JRR to similar methods such as the Balanced Repeated Replication because the JRR is generally simpler and more flexible.

Originally introduced as a technique of bias reduction, the Jackknife method has by now been widely tested and used for variance estimation (Durbin, 1959). Efron and Stein (1981) provide a discussion of the Jackknife methodology. As a landmark empirical study of such applications, see Kish and Frankel (1974). For a general description of JRR and other practical variance estimation methods in large-scale surveys, see Verma (1993).

The JRR variance estimates take into account the effect on variance of aspects of the estimation process which are allowed to vary from one replication to another. In principle this can include complex effects such as those of imputation and weighting. But it has to be noted that often in practice it is not possible to repeat such operations entirely fresh at each replication.

The basic model of the JRR for application in the context described above may be summarised as follows. Consider a design in which two or more primary units have been selected independently from each stratum in the population. Within each primary sampling unit (PSU), subsampling of any complexity may be involved, including weighting of the ultimate units.

In the 'standard' version, each JRR replication can be formed by eliminating one sample PSU from a particular stratum at a time, and increasing the weight of the remaining sample PSU's in that stratum appropriately so as to obtain an alternative but equally valid estimate to that obtained from the full sample.

The above procedure involves creating as many replications as the number of primary units in the sample. The computational work involved is sometimes reduced by reducing the number of replications required. For instance, the PSUs may be grouped within strata, and JRR replications formed by eliminating a whole group of PSUs at a time. This is possible only when the stratum contains several units. Alternatively, or in addition, the groupings of units may cut across strata. It is also possible to define the replications in the standard way ('delete one-PSU at a time Jackknife'), but actually construct and use only a subsample of those.

In the kind of multistage samples encountered in most national household surveys, it is possible to apply the standard JRR method without such grouping of units. However, one common situation in which grouping of units is unavoidable is when the sample or a part of it is a direct sample of ultimate units or of small clusters, so that the number of replications under 'standard' JRR is too large to be practical. Normally, the appropriate procedure to reduce this number would be to form new computational units by forming random groupings of the units within strata. The presence of small and variable-sized PSUs may also require some grouping in practical application of the procedure.

Briefly, the standard JRR involves the following.

Let u be a full-sample estimate of any complexity, and $u_{(hi)}$ be the estimate produced using the same procedure after eliminating primary unit i in stratum h and increasing the weight of the remaining (a_h-1) units in the stratum by an appropriate factor g_h (see below). Let $u_{(h)}$ be the simple average of the $u_{(hi)}$ over the a_h values of i in h. The variance of u is then estimated as:

$$\mathrm{var}(u) = \Sigma_h \left[(1 - f_h) . \frac{a_h - 1}{a_h} . \Sigma_i (u_{(hi)} - u_{(h)})^2 \right]. \qquad (1)$$

A major advantage of a procedure like the JRR is that, under quite general conditions for the application of the procedure, the same and relatively simple variance estimation formula (1) holds for u of any complexity.

A possible variation which may be mentioned is to replace $u_{(h)}$, the simple average of the $u_{(hi)}$ over the a_h replications created from h, by the *full-sample* estimate u:

$$\mathrm{var}(u) = \Sigma_h \left[(1 - f_h) . \frac{a_h - 1}{a_h} . \Sigma_i (u_{(hi)} - u)^2 \right]. \qquad (1')$$

This version tends to provide a 'conservative' estimate of variance, but normally the difference with (1) is small. We have used form (1) in all the illustrations.

Concerning the re-weighting of units retained in a stratum after dropping one unit, normally the factor g_h is taken as (2.a), but for reasons noted below, we propose the form in (2.w):

$$g_h = a_h / (a_h - 1), \qquad (2.a)$$

$$g_h = w_h / (w_h - w_{hi}) \qquad (2.w)$$

where $w_h = \Sigma_i w_{hi}, w_{hi} = \Sigma_j w_{hij}$, the sum of sample weights of ultimate units j in primary selection i.

Note that (2.a) gives the variance of a simple aggregate, while (2.w) gives the corresponding (lower) variance of a mean, or of total as a ratio estimate.

Form (2.w) is used throughout in our illustrations here. This form retains the total weight of the included sample cases unchanged across the replications created – the same total as that for the full sample. With the sample weights scaled such that their sum is equal (or proportional) to some external more reliable population total, population aggregates from the sample can be estimated more efficiently, often with the same precision as proportions or means.

It is interesting to note that the corresponding treatment of simple aggregate versus mean in the Linearisation method is as follows. As noted, the Linearisation approach involves defining a linearised indicative variable such that its variance approximates the variance of the complex statistic concerned. (i) For estimating variance of the total of a quantity y_i, the linearised indicative variable, of course, is y_i itself. (ii) For estimating variance of its mean $\bar{y}$, the linearised variable is $(y_i - \bar{y})$. (See next section for details).

It can be seen that the use of (2.a) in JRR corresponds to (i), and the use of (2.w) to (ii). Most survey statistics of interest are similar in form to ratios, possibly with the added complexity due to the involvement of additional parameters, themselves estimated from

the sample. Hence form (2.w) is the appropriate one. Similarly in the linearisation method, form (ii) – which expresses the linearised variable such that its expected value is zero – is the appropriate one.

Empirical results comparing the performance of the JRR and Linearisation methods in variance estimation of poverty and inequality measures will be presented in Section 5. Firstly, Table 1 shows a few preliminary results of methodological interest, based on a living conditions survey in Toscana.

Table 1. Some illustrations based on a survey in Toscana region

1(a). Alternative JRR variance estimation formulae

	Equivalised income		Poverty rate	
	mean	total	below 50% of mean	below 60% of median
Estimate	34,661	90,428	16.0	17.1
Standard error				
'JRR standard' (1)	5,71	1,489	0.81	0.65
Equation (1')	5,77	1,503	0.83	0.67

1(b). Sampling error of mean income versus that of total income

	'JRR standard' (2.w)		JRR eq. (2.a)		Linearisation	
Estimate	se	%se	se	%se	se	%se
Mean equivalised income						
34,661	571	1.65	566	1.63	564	1.63
Total income ('000)						
90,428	1,489	1.65	2,777	3.07	2,777	3.07

1(c). Effect of treating poverty rate as a simple ratio

Poverty rate	Estimate	'JRR standard'	'JRR fixed'
below 50% of mean	16.0	0.81	0.74
below 60% of median	17.1	0.65	0.79

A few results are shown for mean and total equivalised income, and poverty rates defined in relation to the mean and median incomes. Mean equivalised income is the main variable

of interest in the analysis of poverty and inequality. The income of each household is 'equivalised' using a scale taking into account household size and composition. This equivalised income is then ascribed to each member of the household, and thereafter treated as individual income. The individual forms unit of analysis in the study of income distribution. Conventionally, poverty line is taken as a certain percentage of the mean or the median equivalised income (commonly as 50% of the mean, or 60% of the median). Poverty rate is the proportion of the population with equivalised income below the poverty line.

Table 1(a) shows the difference between forms (1) and (1') of the JRR variance estimation equation. We refer to (1) as the 'standard JRR' version. As noted, (1') provides a more conservative estimate. As seen in the table, the difference from the standard is very small for all the statistics shown. These results are quite typical.

Two further important methodological points may be noted. The first point is of general relevance. With the proposed modification (2.w) to the JRR variance estimator, estimates of mean and total income have exactly the same (relative) precision, as seen in Table 1(b) for 'JRR standard'. This is correct when aggregates are obtained in the form of ratio estimates using fixed external control totals for the base population. Such controls are normally applied in practice in estimating from survey results - hence we use form (2.w) as the norm. Form (2.a) corresponds to simple linear estimate of the aggregate, in which case the JRR variance estimate (3.07%) is found to be practically identical to the Taylor estimate for the same, but much higher than that of the mean (1.65%).

The second point is particularly relevant in the context of analysis of poverty and income inequality. Statistics like the poverty rate are in fact more complex than ordinary ratios, since the threshold defining the attribute (poverty line) is itself subject to sampling variability. This can be taken into account in the JRR method by determining the poverty line separately for each replication. Table 1(c) shows the effect of the treating the poverty rate as a simple proportion as opposed to the more complex statistic defined in terms of an estimated poverty line itself subject to sampling variability. In this particular example, the effect is not large; it is also not found to be in the same direction for the two measures of the poverty rate.

3. Variance estimation based on linearisation

Variance estimation based on Taylor approximation has been widely used and tested, and provides a useful basis of validating other approaches for the type of statistics for which the Taylor approach can be used as an alternative. The basis of the method is the simple variance estimation formula (3) for aggregates in multistage stratified samples of large size.

Let $u_{hi} = u(y_{hi}, x_{hi}, ...)$; $u_h = \Sigma_i u_{hi}$ be a sample aggregate or a linear function of sample aggregates such as of $y = \Sigma_h y_h$; $y_h = \Sigma_i y_{hi}$; $y_{hi} = \Sigma_j (w_{hij}.y_{hij})$. Then its variance is estimated as:

$$var(u) = \Sigma_h \left[(1 - f_h) . \frac{a_h}{a_h - 1} . \Sigma_i \left(u_{hi} - \frac{u_h}{a_h} \right)^2 \right], \qquad (3)$$

with the quantity u_{hi} defined at the level of primary selection (h,i). Here j refers to ultimate sampling unit, i to PSU, and h to stratum; a_h>1 is the number of sample PSU's in stratum h; and $(1-f_h)$ the finite population correction, usually ~1.

3.1 Linearisation procedure

This procedure is extended to non-linear statistics on the basis of Taylor linearisation. The linearisation approach to approximating variance of complex (non-linear) statistics is a long-established procedure; see for instance Deming (1943), Kendall and Stuart (1958), Keyfitz (1957). The basis of the approach is to use Taylor approximation to reduce non-linear statistics to a linear form, justified on the basis of asymptotic properties of large populations and samples. A well-known example is that of a ratio $r = y/x$, for which the linearised indicative variable to be used in equation (3) is $u_{hi} = \frac{y}{x}.\left(\frac{y_{hi}}{y} - \frac{x_{hi}}{x}\right)$.

In more general terms, the basic linearisation procedure may be described as follows. Let $y=(y_1,y_2,y_3,\ldots y_k, \ldots)$ be a vector of statistics with $E(y)=Y=(Y_1,Y_2,Y_3,\ldots Y_k, \ldots)$, and $\lambda(Y)$ a population parameter estimated by $\lambda(y)$. To terms of first degree in (y_k-Y_k), we have $\lambda(y)-\lambda(Y) = \Sigma_k (y_k - Y_k).\left.\frac{\partial\lambda}{\partial Y_k}\right|_y$, so that the required variance of $\lambda(y)$ is estimated by variance of the linearised statistic $\Sigma y_k.\left.\frac{\partial\lambda}{\partial Y_k}\right|_y$, with the partial derivatives evaluated at Y=y. These quantities involve summation over sample units. Woodruff (1971) showed how the computations may be simplified by merely reversing the summation between sample units (i) and component variables (k). The result is to provide, for each sample unit i, a *linearised 'indicative' variable*, say λ_i, such that variance of its aggregate or mean approximates the variance of λ, the complex statistic of interest. The linearisation approach has been further developed and applied extensively in recent years; see for instance the work of Binder and colleagues (Binder, 1983; Binder and Patek, 1994; Binder and Kovacevic 1995; Kovacevic and Yung, 1997), Preston (1995), Deville (1999), Zheng (2001), Demnati and Rao (2004).

The following outlines the procedure we have used to derive the required linearised indicative variables for estimating variances of complex inequality and poverty measures.

1. *Estimation equation and substitution estimator*

Let y_i denote the vector of values of variables y for individual units $i \in U$ in the population, $\lambda = \lambda(y_i \in U, \Lambda)$ the parameter of interest for which the estimation of sampling error is required, and $\Lambda = (\lambda_1, \lambda_2, \ldots)$ a vector of other parameters involved in the definition of λ. For instance, in estimating poverty rate p, defined as the proportion of the population with income y_i below a certain poverty line y_p, we have $\lambda=p$ and a single parameter $\lambda_1=y_p$.

It is convenient to write $\lambda = \lambda(y_i, \Lambda)$ in the form of the *estimation function or equation*:

$$0 = T(y_i \in \cup, \Lambda, \lambda) = \frac{1}{N}.\Sigma_{\cup} T_i \text{ , say,}$$

with the corresponding *substitution estimator* (Rao, 1979):

$$0 = \Sigma w_i.T_i,$$

with Σ over the sample, and:

$$\Sigma w_i = 1,\ w_{i\in s} \neq 0,\ w_{i\notin s} = 0.$$

For example, for poverty rate p defined in relation to a poverty line y_p,

$$T = \Sigma w_i.(\delta(y_i \le y_p) - p) = 0; T_i = \delta(y_i \le y_p) - p; \delta(..) = 1 \text{ if } y_i \le y_p, = 0 \text{ otherwise}.$$

2. *Influence function and the linearised indicative variable*

Under Taylor linearisation, the substitution estimator gives:

$$0 = T_i + \left(\frac{\partial T}{\partial \Lambda}\right)'.\Lambda_1 + \frac{\partial T}{\partial \lambda}.\lambda_i; \text{ where} \left(\frac{\partial T}{\partial \Lambda}\right)'.\Lambda_1 = \frac{\partial T}{\partial \lambda_1}.\lambda_{1i} + \frac{\partial T}{\partial \lambda_2}.\lambda_{2i} + \ldots . \quad (4)$$

Here $(\partial T/\partial \lambda)$ etc. are partial derivatives of T w.r.t λ, and λ_i etc. are the so-called influence functions of the parameters at i. Essentially, an influence function is the derivatives of the parameter concerned at discrete points in the sample space. The important point for the present purpose is that influence function λ_i *is the required linearised indicative variable* for λ in the sense that Var($\Sigma\lambda_i$) approximates Var(λ). (Deville, 1999).

3. An *indicative variable* is specific to the parameter concerned, irrespective of the particular estimation equation in which that parameter appears. Hence once obtained, the indicative variable of a parameter is 'portable' across different estimation equations wherever that parameter is involved.

4. By contrast, the *partial derivatives* are specific to the particular estimation equation T=0 being considered. Ordinary rules of differentiation provide the required derivatives. A few of the most useful rules may be noted.

- Generally, the functional forms of interest are such that the differentiation can be moved under the summation sign:

$$T = \Sigma w_i.T_i, \quad \rightarrow \quad \frac{\partial T}{\partial \lambda} = \Sigma w_i.\frac{\partial T_i}{\partial \lambda} \quad .$$

- A useful case is when the required parameters is a function only of other parameters $\lambda = \lambda(\Lambda)$, without reference to individual element values y_i. In this case:

$$\lambda_i = \left(\frac{\partial \lambda}{\partial \Lambda}\right)'.\Lambda_i = \frac{\partial \lambda}{\partial \lambda_1}.\lambda_{1i} + \frac{\partial \lambda}{\partial \lambda_2}.\lambda_{2i} + \ldots$$

- For a distribution function, evaluated at a particular point y_α:

$F_\alpha = \sum w_i \cdot \delta(y_i \le y_\alpha) = \alpha$, we have as its indicative variable:

$\left(\frac{\partial F}{\partial y}\right)_{y_\alpha} = f_\alpha$, the density function evaluated at that point.

- The general functional forms of interest in the context of poverty and inequality measures is $\lambda_\alpha = \sum w_i \cdot k(y_\alpha) \cdot \delta(y_i \le y_\alpha)$, giving:

$$\frac{\partial \lambda}{\partial y_\alpha} = \sum w_i \cdot \left(\frac{\partial k}{\partial y_\alpha}\right) \cdot \delta(y_i \le y_\alpha) + k_\alpha \cdot f_\alpha .$$

Two forms of the above are of interest:

i) k is a constant or does not involve y_α, so that only the second term above is present. This is the density function as noted above.

ii) k is a function of y_α, but such that k_α, its value at α, is zero. In this case only the first term in the above equation applies.

5. The linearised variables for poverty measures involve reference to the *density function* at various points in income distribution, such as at the median or the poverty line. Special procedures are required in density function estimation, as described in Section 3.2 below.

Using these rules, we have derived the required linearised variables for a full range of poverty and inequality measures. These have been listed in Annex I for future reference.

Note that, on the basis of equation (4) we have divided each linearised variable λ_i into two parts (each with zero sample mean):

$$\lambda_i = \left(\frac{T_i}{-\partial T/\partial\lambda}\right) + \frac{1}{-\partial T/\partial\lambda}.\left(\frac{\partial T}{\partial\lambda_1}.\lambda_{1i} + \frac{\partial T}{\partial\lambda_2}.\lambda_{2i} + \ldots\right) = \lambda_i^{(0)} + \lambda_i^{(v)} \text{ , say.} \qquad (5)$$

Using only the first term of (5) as the linearised variable gives the estimate of var(λ) with parameters Λ treated as constants, i.e, with λ treated as a simple aggregate or ratio. Including both parts incorporates the effect of sample dependence of parameters Λ involved in the definition of λ. Treating these parameters as constants will of course greatly simplify the estimation of sampling error of the measure concerned. It is therefore of practical interest to identify separately the extent to which such a simple but crude estimate is modified as a result of sampling variability of the parameters involved. Consider for instance a measure such as the proportion poor in a population. It is more complex than an ordinary proportion, in so far as it is defined in relation to some poverty line (such as a given fraction of mean or median income) which is itself subject to sampling variability. How large is the effect of this complexity on the magnitude of sampling error of the resulting statistic? Specifically, can it be acceptable to treat the poverty rate simply as an ordinary proportion in estimating its sampling error? This issue arises in relation to other measures as well. We will provide a number of numerical illustrations below.

3.2 Density function estimation

A brief note follows on the methodology of density function estimation required in the application of the linearisation variance estimation methodology.

Even though the techniques of nonparametric density estimation have a long tradition, new theories and methodologies have been developed and expanded only in the last two decades. The growing interest is owed essentially to two factors: first of all, scholars of statistics have found that purely parametric estimation of curves is not always sufficiently flexible for data analysis. Moreover, the development of computing power has cleared the path for nonparametric estimates that were not feasible in the past.

It will be useful to introduce here the basic idea of nonparametric density estimation. Let $\{y_i\}_{i=1}^n$ be the values of the variable of interest, observed on a set of n units. The aim in the linearisation method is to estimate the derivative of the distribution function at a certain point y_α: $f_\alpha = f(y_\alpha) = (dF/dy)y_\alpha$, which is the density function $f(y)$ of y at y_α, estimated as $\hat{f}(y_\alpha)$. The nonparametric density function estimator is defined as a local average of the observations found in a band around the point (y) at which the value is to be estimated:

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} w_i(y,\mathbf{y}) \ , \tag{6}$$

where $\{w_i(y,\mathbf{y})\}_{i=1}^{n}$ denotes a sequence of values that depends on the vector $\mathbf{y}$ containing the values $\{y_i\}_{i=1}^{n}$, and the value y at which the function is estimated. This estimator is defined as *smoothing*, while the estimate is called *smoother*.

Among the most important smoothing techniques (the manner in which succession of weights is calculated) one can list the kernel, the k_{th} closest point, the orthogonal series and the 'spline smoothing'. The most utilised of these is the kernel technique, adopted also in the present illustrations.

In kernel smoothing the sequence of values is defined as:

$$w_i = \frac{1}{h}K\left(\frac{y - y_i}{h}\right) . \tag{7}$$

Here K(.) is the kernel, a symmetric, limited, continuous function whose integral is equal to one on the interval for which it is defined; h is the bandwidth or *smoothing parameter*.

This parameter regulates the width of the interval around y. A local average for an interval too wide can lead to the consideration of observations that have little in common with y. On the other hand, consideration of a low number of observations can make the estimate $f(y)$ too irregular and can inflate the variability greatly.

The shape of the kernel function regulates the way in which values diminish as we move away from y. Substituting the formula (7) into the smoothing (6) one gets:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{y - y_i}{h}\right). \tag{8}$$

The choice of the kernel function and the band parameter is intended to minimise the distortion and variability of the estimate of the function f(y). For this purpose one precision measure of global accuracy is considered: the mean integrated squared error (MISE), which is defined as:

$$MISE(\hat{f}) = E\int\{\hat{f}(y) - f(y)\}^2 dy .$$

Under the simple assumptions contained in Silverman (1986) and the additional assumption that the function $f(y)$ has continuous derivatives of at least the second order, one can approximate the bias and the variance of $\hat{f}(y)$ and thus calculate the MISE:

$$MISE(\hat{f}(y)) = \frac{1}{4}h^4 k_2{}^2 \int_{D(x)} f''(y)^2 dy + \frac{1}{nh}\int_{-\infty}^{+\infty} K(t)^2 dt , \tag{9}$$

where $k_2 = \int_{-\infty}^{+\infty} t^2 K(t)dt$, and $f''(y)$ is the second order derivative of $f(y)$. The optimal value of h derived by minimisation of (9) is:

$$h = k_2^{-2/5}\left\{\int_{-\infty}^{+\infty} K(t)^2\,dt\right\}^{1/5}\left\{\int_{D(y)} f''(y)^2\,dy\right\}^{-1/5} n^{-1/5}. \qquad (10)$$

This value cannot be calculated for the unknown function f(y); the way to obtain it is shown below.

Substituting the previous formula in the equation for MISE one gets: $MISE = (5/4)C(K)\{\int f''(y)^2 dy\}^{1/5} n^{-4/5}$, where the function C(K) is :

$$C(K) = k_2^{2/5}\{\int K(t)^2 dt\}^{4/5}.$$

By analysing this formula, we observe that the kernel function minimising the MISE, holding other parameters constant, is the same as the one minimising the function C(K). On the basis of this kernel the efficiency of a generic kernel is defined as $eff(K) = \{C(K_e)/C(K)\}^{5/4}$. The efficiency of the most common kernels is almost equal to one; the consequence is that the choice of kernel is not crucial for the convergence of the estimate $\hat{f}(y)$ to $f(y)$, while the bandwidth h parameter is. In the present illustrations, the kernel with a normal distribution has been utilised for its analytical properties. The study of the minimisation of the MISE is thus focused on the optimisation of the smoothing parameter, which however cannot be calculated directly from (5). In the choice of the smoothing parameter there is a trade-off between the mean and the variance of $\hat{f}(y)$; a value for h too small makes the estimate jagged, with minimum distortion but high variance. By contrast, a high value of h makes the estimate homogenous, with low variance but strong distortion.

Among the more common techniques for choosing the parameter h is the 'plug-in method'. The functional form of $f(y)$ is hypothesised *a priori*, from which the second derivative is substituted in (10). When f represents an income density function and it is expected to come from a log-normal or heavily skewed distribution, an optimal value of the bandwidth parameter can be evaluated as $h_{opt} = 0.79Rn^{-1/5}$ (Silverman, 1986, p.47), where R is the interquantile range.

3.3 Sensitivity analysis of density function estimation

With the linearisation approach, a potential source of uncertainty arises from the need to estimate density function of the income distribution from numerical data, which are generally subject – as in the data used for the present illustrations – to significant irregularities and 'lumpiness'. The numerical estimates of variance depend on how the density function is evaluated, specifically the degree of smoothing applied to the numerical data.

In the following examples of sensitivity analysis[1] we have used a bandwidth parameter for the smoothing of the density function calculated with the plug-in method as $h_{opt} = 0.79Rn^{-1/5}$, where R is the interquantile range, corresponding in this example to

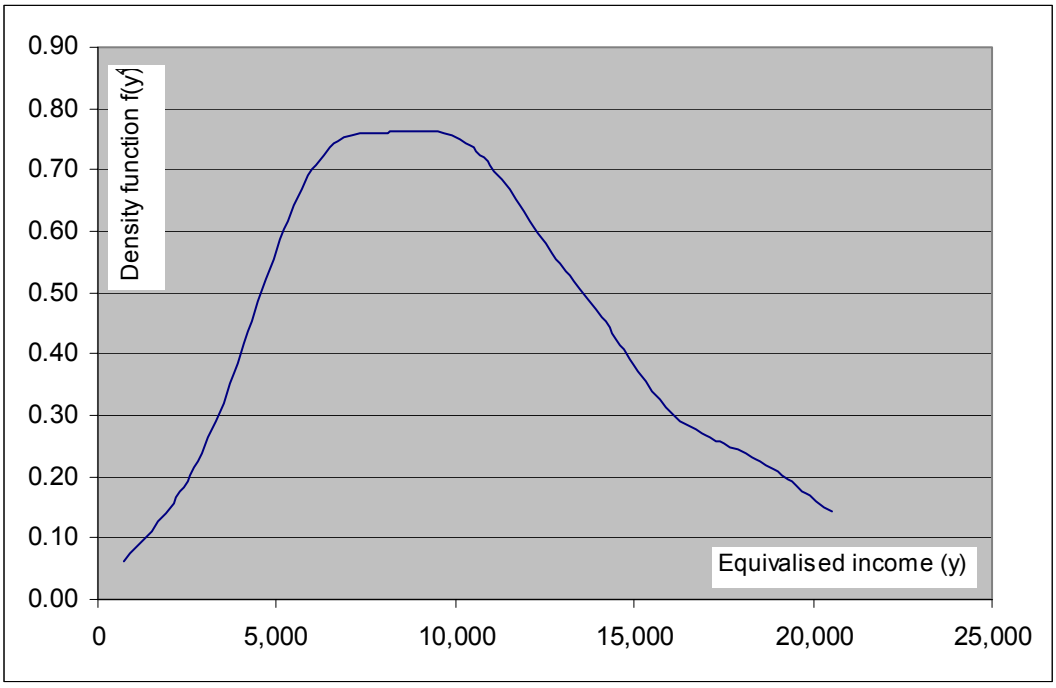
[1] These examples are based on some data from the European Community Household Panel (ECHP). More detailed results using these data are presented in Section 5 below.

about 1,066 monetary units in the equivalised income. The density function shape and some values corresponding to various selected values of the income distribution are reported in Figure 1.

In order to validate these results we have performed a sensitive analysis of the choice of the bandwidth parameter.

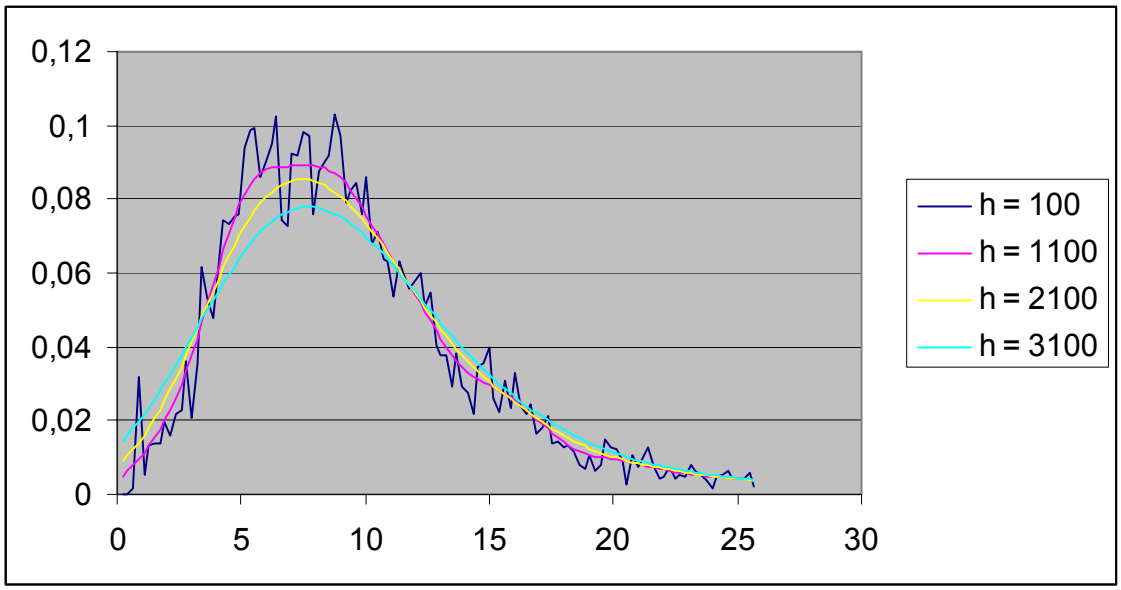
When performing nonparametric density estimation a crucial issue is the choice of the bandwidth parameter h. A range of values is possible depending on which method and which formula we start with. In order to show how sensitive can the estimated density function be to the bandwidth parameter, in Figure 2 we report four distributions estimated with parameters ranging from 100 monetary units (small bandwidth) to more than 3,000 monetary units (large bandwidth). The distribution with the smallest bandwidth is characterised by a spuriously fine structure and it is very sensitive to irregularities in the values of $\{y_i\}_{i=1}^{n}$ from the sample. As we increase the value of the bandwidth, the density function becomes smoother and more homogenous. However, with values too large the central part of the distribution is clearly underestimated , while the tails are overestimated since the estimate is affected by points (those in the central part of the distribution) quite far away from y, the point of interest.

Figure 1. Density function estimation



percentile	equivalised income (y)	density function $f(y)*10^4$
P10	4.866	0,5484
P20	6.287	0,7236
P50 (median)	10.178	0,7486
P80	15.564	0,3358
P90	19.556	0,1837
60%median	6.107	0,7095
50%mean	5.841	0,6838
Band width, h=	1.066	

Figure 2. Sensitivity analysis over a wide range of h values



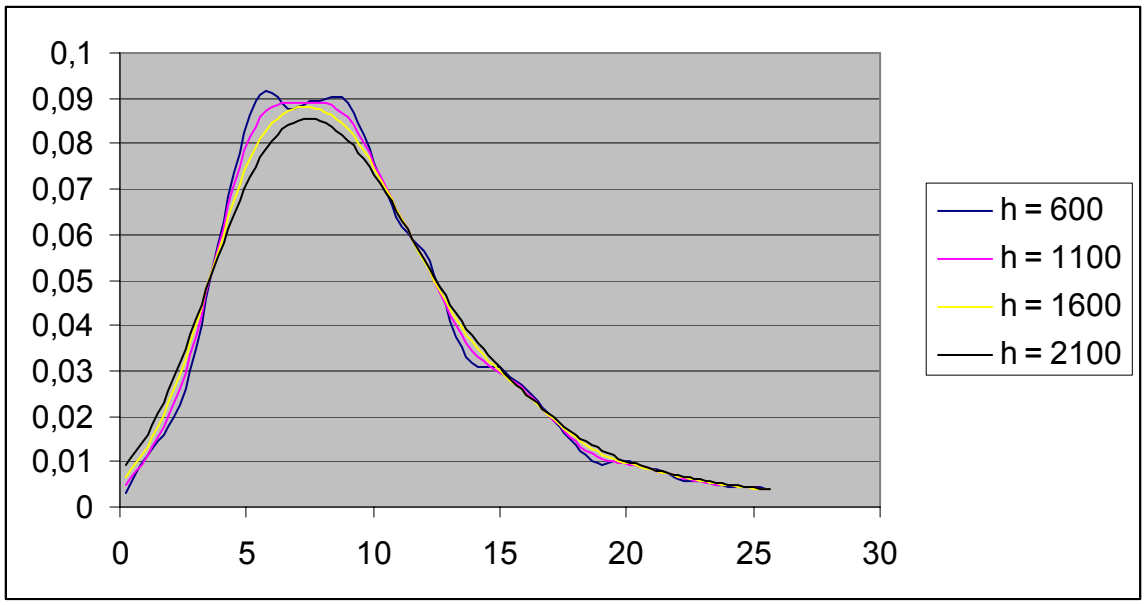
Going back to the methods and formulas for choosing the bandwidth parameter, it is important to underline that most of the common methods, i.e. least-squares cross-validation, likelihood cross-validation, the test graph method, etc. are based on some assumptions which analysts involved in applications are rarely able to check.

Even the simpler plug-in method proposed here is based on some assumptions about the functional form of the density function f. The optimal value of the bandwidth parameter depends on that form.

For instance, when the function is symmetric with a kurtosis similar to the Gaussian distribution, the optimal value of the bandwidth parameter can be approximated by $h_{opt} = 1.06\sigma n^{-1/5}$; when the distribution is unimodal, but asymmetric and with a kurtosis dissimilar from the Gaussian distribution (e.g. lognormal or t-family distributions) the optimal value of the bandwidth parameter can be approximated by $h_{opt} = 0.79Rn^{-1/5}$; finally in the case of bimodal distributions (for instance a mixture of income distributions in a very polarized society or in a country with very marked regional differences) the optimal value of the bandwidth parameter can be approximated by $h_{opt} = 0.9An^{-1/5}$ where $A = \min(\sigma, R/1.34)$ (Silverman, 1986, p. 47).

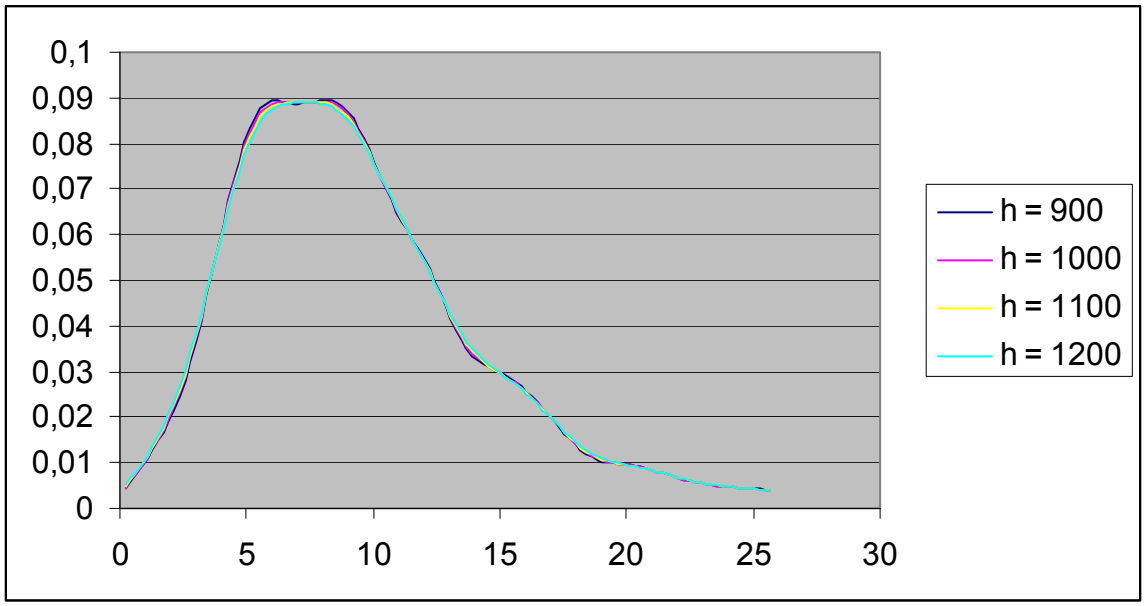
These three different formulas of the same method for calculating the bandwidth parameter lead to three quite different values (h=905, h=1066 and h=2329), any of which is clearly admissible when the original shape of the distribution is unknown. Figure 3 shows again four estimated density function based on four admissible parameters ranging from 600 to 2100 monetary units. The distribution show some differences, particularly evident in the central part of the distribution where are located the mean and the median, two fundamental points at which density needs to be calculated for variance estimation of poverty measures.

Figure 3. Sensitivity analysis over a medium range of h values



Finally the last set of distributions reported in Figure 4 show that the distributions and the results are not particularly sensitive to the choice of the value h only when the range of the bandwidth parameter is limited.

Figure 4. Sensitivity analysis over a limited range of h values



4. Structure of the variance computation algorithm

In this section we describe the basic structure of the computational algorithms of the JRR and the Linearisation procedures, as we have implemented those.

It is worth noting again that measures such as the proportion poor in a population are more complex than ordinary proportions, in so far as the former are defined in relation to some poverty line (such as a given fraction of mean or median income) itself subject to sampling variability. How large is the effect of this complexity on the magnitude of sampling error of the resulting statistic? This issue arises in relation to other measures as well. For instance, the mean income of an income quantile or the income share of the quantile (the Lorenz curve ordinate) may be treated as an ordinary mean or ratio, but it also depends on the quantile boundary which is subject to sampling variability.

More generally, most measures of poverty and inequality can be written (and computed) in the form of ordinary ratios, but involving parameters determined from the sample. It is of practical interest to identify separately the extent to which such a simple but crude estimate is modified as a result of sampling variability of the parameters involved.

4.1 Structure of the variance computation algorithm: JRR

In view of the above, we have implemented JRR in two versions:

(1) In the simplified version, we write the poverty or inequality measure of interest in the form of an ordinary aggregate or ratio, treating the parameters involved in the definition of the measures as constants. In other words, these parameters are computed only once based on micro data for the full sample and are used unchanged in each replication.

(2) The actual results are produced by treating the parameters involved as variable from one replication to another. In other words, for each replication, the parameters are recomputed based on micro data for the weighted sample cases included in that replication.

In either case, we begin by defining the parameters (Λ) and variable of interest at the micro-level (u_i) for the full sample, and construct the replications (k) using the standard JRR approach. Note that u_i is a function of the unit concerned (i), and the aggregate parameters Λ are estimated from the sample.

In the 'constant parameter' version, the micro-level variables (u_i), as defined using the parameters Λ estimated from the full sample, are used unchanged, the only difference being the set of units and their adjusted weights included in the computation of the required statistic (say U_k) for each replication k.

In the 'variable parameter' (i.e., real) version, the micro-level variables (u_i) are redefined in each replication, using the parameters Λ estimated for that replication.

Table below illustrates the difference between procedures (1) and (2) for JRR.

Full sample

$$\Lambda = \Lambda(s); \quad u_i = u(s_i, \Lambda) \quad U = \Sigma_{i \in s} w_i . u_i$$

Construction of jackknife replications

Replication	(1) CONSTANT PARAS	(2) VARIABLE PARAMETERS		
	Replication aggregate	Parameters	Unit values	Replication aggregate
S_1	$U_1 = \Sigma_{i \in 1}(w_{i,1}.u_i)$	$\Lambda_1 = \Lambda(S_1)$	$u_{i,1} = u(s_i, \Lambda_1)$	$U_1 = \Sigma_{i \in 1}(w_{i,1}.u_{i,1})$
….	….	….	….	….
S_k	$U_k = \Sigma_{i \in k}(w_{i,k}.u_i)$	$\Lambda_k = \Lambda(S_k)$	$u_{i,k} = u(s_i, \Lambda_k)$	$U_k = \Sigma_{i \in k}(w_{i,k}.u_{i,k})$
….	….	….	….	….
S_K	$U_K = \Sigma_{i \in K}(w_{i,K}.u_i)$	$\Lambda_K = \Lambda(S_K)$	$u_{i,K} = u(s_i, \Lambda_K)$	$U_K = \Sigma_{i \in K}(w_{i,K}.u_{i,K})$

The following notation has been used. S_k refers to replication k; s_i are the values for a variable or set of variables for unit i in the sample; u_i refers to the variable for unit i, the weighted sum of which gives the statistic of interest U. $u_{i,k}$ and U_k refers to the corresponding quantities for a particular replication k. Λ is the set of parameters, estimated form the sample, which are involved in the definition of U and u_i. Λ_k is the corresponding estimate based on replication k.

Sample weights

Other aspects, such as 'raking' of sample weights to match external total of auxiliary variables may be incorporated by redetermining the weights in each replication using the same procedure.

4.2 Structure of the variance estimation algorithm: Linearisation

The following shows how exactly the same distinction is implemented in the linearisation approach.

Full sample

Linearisation method, of course, always refers only to the full sample. The basic quantities are computed as in the JRR case.

$$\Lambda = \Lambda(s); \quad u_i = u(s_i, \Lambda) \quad u = \Sigma_{i \in s} w_i . u_i$$

The difference between the simplified 'constant parameter' and the actual 'variable parameter' version arise from how the linearised indicative variables u_i are defined.

As noted in section 3.1, the linearised variable u_i may be divided into two parts, say:

$$u_i = u_i^{(0)} + u_i^{(v)} \tag{11}$$

where the first part corresponds to the required linearised variable when the statistic of interest is treated as if it were a simple ratio (i.e. the parameters involved in its definition taken as constants).

Constant parameter version

- take $u_i^{(0)}$ as the linearised variable for variance estimation

Variable parameter version

- take the full expression (11) as the linearised variable

Sample weights

Special procedures are required to accommodate 'raking' of sample weights to match external totals of auxiliary variables. The procedure would generally differ depending on whether or not the auxiliary variables are available at the micro-level.

5. Illustrative results

5.1 Main results

Before discussing details, Table 2 shows some main results from a national survey based on a complex (stratified, two-stage, weighted) sample. The sample consisted of 250 clusters selected systematically with probability proportional to size sampling. Each selected cluster was subsampled with probability inversely proportional to cluster size to obtain a sample of 4,800 households, containing nearly 12,000 persons in total. Complex procedures were used to impute missing data and compute unit weights. Throughout we take these imputations and sample weights as given constants.

For the purpose of computing sampling errors, these 250 sample clusters were paired in the order of selection to obtain 125 computing strata.

The table shows standard errors for the actual design, and also what they would be for a simple random sample (SRS) of the same size. Ratio of actual to SRS standard error, the design effect or 'deft', is the factor by which the width of the confidence interval is inflated due to departures of the design from SRS.

The variables referred to in the table are well-known measures used in the analysis of poverty and income distribution. Please see Annex I for their definition in a weighted sample.

A notable feature of the results in Table 2 is that, for the same design, design effects can differ greatly from one type of variable to another: there is no one value of *the* design effect for a given sample design. This feature has been amply confirmed in other large scale multi-country studies of design effects (e.g., Verma, Scott and O'Muircheartaigh, 1980; Verma and Lê, 1996).

Table 2
JRR estimates of sampling errors and design effects for measures of poverty and income distribution

	Measure		Estimate [1]	standard error [2]	relative error (%) actual [3]	relative error (%) SRS [4]	deft [3]/[4] [5]
1	Mean Equivalent Income		11,681	177	1.5	0.9	1.6
2	percentile P10		4,866	73	1.5	0.6	2.5
3	percentile P20		6,287	116	1.8	1.4	1.4
4	Income median (P50)		10,177	216	2.1	1.4	1.5
5	percentile P80		15,564	283	1.8	1.0	1.8
6	percentile P90		19,555	294	1.5	1.1	1.4
7	HCR based on 60% median		18.51	0.63	3.4	3.0	1.1
8	HCR based on 50% mean		16.31	0.72	4.4	4.0	1.1
9	FGT, e=1 (poverty gap)		5.22	0.34	6.5	4.7	1.4
10	FGT, e=2		2.38	0.23	9.5	6.2	1.5
11	SEN		7.46	0.44	5.9	4.2	1.4
12	RMD		21.85	0.38	1.7	1.4	1.3
13	Theil		17.21	0.65	3.8	4.0	1.0
14	Atkinson e=1		15.79	0.50	3.2	2.7	1.2
15	Gini		31.15	0.51	1.6	1.4	1.2
16	percentile ratio P80/P20		2.48	0.05	2.0	1.6	1.3
17	percentile ratio P90/P10		4.02	0.08	1.9	1.2	1.6
18	Mean Equivalent Income (S10)		3,427	136	4.0	2.5	1.6
19	Mean Equivalent Income (S20)		4,520	111	2.4	1.6	1.5
20	Mean Equivalent Income (S80)		22,671	469	2.1	1.6	1.3
21	Mean Equivalent Income (S90)		27,894	715	2.6	2.5	1.0
22	%share S10		2.91	0.11	3.6	3.5	1.0
23	%share S20		7.73	0.17	2.2	2.0	1.1
24	%share S80		39.06	0.45	1.1	1.1	1.0
25	%share S90		24.10	0.43	1.8	1.7	1.1
26	share ratio S80/S20		5.05	0.15	2.9	2.7	1.1
27	share ratio S90/S10		8.27	0.35	4.3	4.2	1.0
Of all variables 01-27		average			3.0	2.4	1.3
		min			1.1	0.6	1.0
		max			9.5	6.2	2.5
Of variables 01-15		average			3.4	2.5	1.4
		min			1.5	0.6	1.0
		max			9.5	6.2	2.5
Excluding 1 extreme value at each end							
Of all variables 01-27		average			2.8	2.3	1.3
		min			1.5	0.9	1.0
		max			6.5	4.7	1.8
Of variables 01-15		**average**			2.9	2.4	1.4
		min			1.5	0.9	1.1
		max			6.5	4.7	1.8

5.2 Comparison with the linearisation approach

Table 3 compares the results using the linearisation approach with those from the previous table using JRR. The general closeness of the results from these two entirely different methodologies is, in our view, quite remarkable. There are some significant differences

however, particularly concerning the quantiles, especially the lowest decile (P10). The difference for the median and the highest decile is also of the order of 15-20%. Limitation of the JRR method for estimating variance of quantiles has been noted in the literature. With the linearisation approach, a potential source of uncertainty arises from the need to estimate density function of the income distribution from numerical data which are generally subject – as in the present case – to significant irregularities and 'lumpiness'. The numerical estimates of variance depend on how the density function is evaluated, specifically the degree of smoothing applied to numerical data, as discussed in Section 3.3.

Table 3

Estimates of relative standard error: JRR and Taylor linearisation comparison

		estimate	%standard error		ratio
			JRR	Taylor	Taylor/JRR
		[1]	[2]	[3]	[4]
1	Mean Equivalent Income (all)	11,681	1.51	1.57	1.04
2	percentile P10	4,866	1.50	2.49	1.66
3	percentile P20	6,287	1.84	1.97	1.07
4	Income median (P50)	10,177	2.12	1.77	0.83
5	percentile P80	15,564	1.82	1.88	1.03
6	percentile P90	19,555	1.51	1.80	1.19
7	HCR based on 60% median	18.51	3.43	3.56	1.04
8	HCR based on 50% mean	16.31	4.40	4.13	0.94
9	FGT, e=1 (poverty gap)	5.22	6.53	5.64	0.86
10	FGT, e=2	2.38	9.46	7.94	0.84
11	SEN	7.46	5.90	6.67	1.13
12	RMD	21.85	1.72	1.64	0.95
13	Theil	17.21	3.80	3.54	0.93
14	Atkinson e=1	15.79	3.19	2.80	0.88
15	Gini	31.15	1.63	1.67	1.02
	average		3.36	3.31	1.03
	min		1.50	1.57	0.83
	max		9.46	7.94	1.66

5.3 Effect of treating a complex statistic as a simple ratio

In Table 4, the first four columns show variance estimates obtained by treating each complex statistic as a simple ratio. As explained earlier, for JRR this implies that any sample dependent parameters involved are determined only once, from the total sample, rather than repeatedly from each replication. For the linearisation method, this means using only the first of the two parts of the linearised variable for the parameter concerned as defined in Annex I.

No estimates of this type can be produced for quantiles of the distribution, such as median income. For the other variables shown in Table 4, the agreement between the two methods is again generally very close (see the ratio in column [4]).

Table 4
Estimates of relative standard error: treating estimates as simple ratios*

	estimate	%standard error		ratio	Ratio to corresponding cols. of Table 3		
		JRR	Taylor	Tay/JRR	JRR	Taylor	Tay/JRR
	[1]	[2]	[3]	[4]	col[2]	col[3]	col[4]
1 Mean Equivalent Income (all)	11,681	1.51	1.57	1.04	1.00	1.00	1.00
2 percentile P10	4,866						
3 percentile P20	6,287						
4 Income median (P50)	10,177						
5 percentile P80	15,564						
6 percentile P90	19,555						
7 HCR based on 60% median	18.51	4.47	4.71	1.05	1.30	1.32	1.01
8 HCR based on 50% mean	16.31	4.82	5.21	1.08	1.09	1.26	1.15
9 FGT, e=1 (poverty gap)	5.22	6.57	5.98	0.91	1.01	1.06	1.05
10 FGT, e=2	2.38	9.45	7.96	0.84	1.00	1.00	1.00
11 SEN	7.46	8.12	7.05	0.87	1.38	1.06	0.77
12 RMD	21.85	2.00	1.94	0.97	1.16	1.18	1.02
13 Theil	17.21	11.61	11.92	1.03	3.05	3.37	1.10
14 Atkinson e=1	15.79	8.25	8.30	1.01	2.59	2.96	1.15
15 Gini	31.15	6.89	7.07	1.03	4.23	3.18	0.75
average		6.37	6.17	0.98	1.78	1.74	1.00
min		1.51	1.57	0.84	1.00	1.00	0.75
max		11.61	11.92	1.08	4.23	3.37	1.15

The last three columns of the table show the comparison of these results with 'proper' variance estimates for the actual complex statistics in Table 3. The remarkable feature of the results is that for a number of measures, *treating them as simple aggregates or means grossly over-estimates the variance*, at least for the survey data in our example. This effect is smaller for poverty measures, but tends to be much larger for the inequality measures considered. The pattern is quite consistent for the two methods, JRR and linearisation, though the actual magnitudes of the effects vary somewhat as seen from the last column of Table 4.

5.4 Design effects

Estimation of the design effect (deft) requires estimation of what the variance would be for the same statistic under simple random sampling (SRS) of the same size. Application of the JRR as defined above does not yield this information directly. We have proposed and used the following procedure for estimating design effect in this situation.

The ratio (i)/(ii) of the quantities defined below estimates SRS error, which forms denominator of deft.

(i) Standard error under random groupings of elements

Replications are constructed as in the normal application of the JRR, but in place of the actual primary selections, *random grouping* of the sample elements are used for this purpose. This provides a variance estimate corresponding to a sample of elements (i.e., without stratification or clustering), but which still differs from the SRS estimate due to the effect of sample weights on variance.

Numerically, the results can be affected by exactly how the replications are formed. On the basis of experience, we recommend that the groups formed should be of *uniform weighted size,* that is, as far as possible, constructed by including a constant sum of weights of elementary units in every grouping. (In practice, we divided the sample of households at random into the same number of groupings as the original number of PSU's, i.e., 250, each

group consisting of a random subset of household with approximately the same total weight).

(ii) Effect of sample weights on standard error

When the sample weights are essentially random (unrelated to unit characteristics), the effect of the weights on increasing the variance is well approximated by the expression:

$$D_W^2 = 1 + cv^2(w_j) \tag{12}$$

where cv is the coefficient of variation of the unit weights. (For our data, this factor equals 1.19.) We may estimate the effect of weighting more precisely, from the following. For statistics such as ratios the following expressions give SRS variance and the same including the inflation due to the effect of weighting:

$$var_{SRS}(u) = \frac{1}{n.(n-1)}.\Sigma_j\left(\frac{w_j}{\overline{w}}\right).u_j^2,\ var_{wt}(u) = \frac{1}{n.(n-1)}.\Sigma_j\left(\frac{w_j}{\overline{w}}\right)^2.u_j^2 \tag{13}$$

with $D_W^2(u) = var_{wt}(u)/var_{SRS}(u)$ estimating the effect of sample weights (specific to each statistics u).

With the linearisation approach, the above expressions can be applied to obtain the SRS error and the effect of weights for more complex statistics for which the required linearised variables have been obtained. Here u_j refers to the linearised indicative variable for the statistic concerned (see Annex 1).

However, with the JRR approach, it cannot be assumed that the full expression for u_j, which comes from the linearised approach, is available. (For one thing, we may be applying the JRR approach for statistics and sample designs for which the linearisation approach is not available). Nevertheless, the first part of the expression for u_j – which corresponds to treating the complex statistics as a simple ratio – is always available (see Annex I).

As an approximation, we may use only this first part for u_j in equation (13) to estimate the effect of weighting. Generally, the approximation is expected to be an improvement over (12).

Table 5 shows the results for the two methods, JRR and Linearisation, treating the sample as a random grouping of elements. These parallel the results in Table 3, except that here the randomised rather than the original clustered sample has been used. (In both cases, the computations are for the actual complex statistics.) The variances correspond to random sample of households (with no clustering or stratification), which differs from a SRS only due to unequal weights.

Generally, the agreement between the two methods remains close - or appears to be even closer with the sample structure randomised in the above sense – for the poverty and in particular the inequality measures. Large differences remain for some of the quantiles, as before. Table 6 shows the results with the above computations repeated with the added assumption of treating the statistics as simple ratios in the sense explained earlier. The main point of these figures is to show how close the results with the two approaches become in most cases.

Finally, Table 7 shows estimates of the effect of weights and hence of the final design effects. In most cases, the effect of weights is predicted reasonably by the simple expression for D_W given in (12), thought form (13) is preferable, in general.

Table 5 ('randomised' sample of households*)
Estimates of relative standard error: JRR and linearisation comparison

		estimate	%standard error		ratio
			JRR	Taylor	Tay/JRR
		[1]	[2]	[3]	[4]
1	Mean Equivalent Income (all)	11,681	1.10	1.10	1.00
2	percentile P10	4,866	0.71	2.01	2.82
3	percentile P20	6,287	1.62	1.60	0.99
4	Income median (P50)	10,177	1.66	1.08	0.65
5	percentile P80	15,564	1.22	1.21	1.00
6	percentile P90	19,555	1.27	1.28	1.01
7	HCR based on 60% median	18.51	3.59	3.47	0.97
8	HCR based on 50% mean	16.31	4.75	4.12	0.87
9	FGT, e=1 (poverty gap)	5.22	5.61	4.89	0.87
10	FGT, e=2	2.38	7.36	6.88	0.93
11	SEN	7.46	4.98	5.90	1.18
12	RMD	21.85	1.64	1.64	1.00
13	Theil	17.21	4.74	4.73	1.00
14	Atkinson e=1	15.79	3.20	3.19	1.00
15	Gini	31.15	1.67	1.67	1.00
	average		3.01	3.02	1.00
	min		0.71	1.08	0.65
	max		7.36	6.88	2.82

* Original sample of households 'randomised'; i.e., retaining original household weights,
the sample households assigned at random to 250 clusters of equal (weighted) size.

Table 6 ('randomised' sample of households*)
Estimates of relative standard error: treating estimates as simple ratios*

		estimate	%standard error		ratio	Ratio to corresponding cols. of Table 3		
			JRR	Taylor	Tay/JRR	JRR	Taylor	Tay/JRR
		[1]	[2]	[3]	[4]	col[2]	col[3]	col[4]
1	Mean Equivalent Income (all)	11,681	1.10	1.10	1.00	1.00	1.00	1.00
2	percentile P10	4,866						
3	percentile P20	6,287						
4	Income median (P50)	10,177						
5	percentile P80	15,564						
6	percentile P90	19,555						
7	HCR based on 60% median	18.51	3.95	3.95	1.00	1.10	1.14	1.03
8	HCR based on 50% mean	16.31	4.41	4.41	1.00	0.93	1.07	1.15
9	FGT, e=1 (poverty gap)	5.22	5.07	5.07	1.00	0.90	1.04	1.15
10	FGT, e=2	2.38	7.03	7.03	1.00	0.95	1.02	1.07
11	SEN	7.46	6.11	6.11	1.00	1.23	1.04	0.84
12	RMD	21.85	1.94	1.94	1.00	1.19	1.18	1.00
13	Theil	17.21	10.78	10.78	1.00	2.27	2.28	1.00
14	Atkinson e=1	15.79	5.37	5.37	1.00	1.68	1.68	1.00
15	Gini	31.15	5.77	5.77	1.00	3.44	2.60	0.75
	average		5.15	5.15	1.00	1.47	1.40	1.00
	min		1.10	1.10	1.00	0.90	1.00	0.75
	max		10.78	10.78	1.00	3.44	2.60	1.15

* All estimates expressed in the form of simple ratios, with any parameters involved treated as constants.
Original sample of households 'randomised'; i.e., retaining original household weights,

Table 7
Effect of sample weights and the estimates of overall design effects

		Estimated effect of sample weights on deft					Estimated design effect (deft)**			
		randomised sample	wtd sample of households		SRS of households	effect of weighting*	Actual estimates (Tables 3 and 5)		treating estimates as simple ratios (Tables 4 and 6)	
		[1]	[2]	ratio [2]/[1]	[3]	ratio [2] / [3]	JRR	Taylor	JRR	Taylor
	Estimate	%standard error					design effect (actual to SRS standard error)			
1	Mean Equivalent Income (all)	11,681	1.10	1.11	1.01	0.97	1.15	1.64	1.70	1.64	1.70
2	percentile P10	4,866						2.51	1.48		
3	percentile P20	6,287						1.36	1.47		
4	Income median (P50)	10,177						1.52	1.95		
5	percentile P80	15,564						1.79	1.85		
6	percentile P90	19,555						1.42	1.67		
7	HCR based on 60% median	18.51	3.95	3.73	0.95	2.98	1.25	1.14	1.22	1.35	1.42
8	HCR based on 50% mean	16.31	4.41	4.17	0.95	3.29	1.27	1.11	1.20	1.31	1.41
9	FGT, e=1 (poverty gap)	5.22	5.07	4.92	0.97	3.95	1.24	1.39	1.38	1.55	1.41
10	FGT, e=2	2.38	7.03	6.71	0.95	5.42	1.24	1.53	1.38	1.61	1.35
11	SEN	7.46	6.11	5.86	0.96	4.72	1.24	1.41	1.35	1.59	1.38
12	RMD	21.85	1.94	1.89	0.98	1.67	1.13	1.26	1.20	1.23	1.20
13	Theil	17.21	10.78	10.44	0.97	9.20	1.13	0.96	0.89	1.29	1.32
14	Atkinson e=1	15.79	5.37	5.49	1.02	4.60	1.19	1.19	1.05	1.84	1.85
15	Gini	31.15	5.77	5.70	0.99	5.06	1.13	1.16	1.20	1.43	1.46
	average (vars 1-15)							1.43	1.40		
	average (vars 1, 7-15)				0.97		**1.20**	1.28	1.26	1.48	1.45
	min				0.95		1.13	0.96	0.89	1.23	1.20
	max				1.02		1.27	2.51	1.95	1.84	1.85

* The average value of this factor is taken as the overall (uniform) effect of sample weights and in inflating standard errors or deft.
This is very close to the 'Kish Factor'=1.19, approximating the effect of 'haphazard' weights (see text).
** Computed as the ratio of standard errors for the actual sample and the corresponding randomised sample of households, multiplied by the average effect of weighting on de

Annex I
Linearised 'indicative' variables for variance estimation: Poverty, inequality and income distribution measures

The following table lists the linearised variables required for the application of the Linearisation method of variance estimation. Expressions are provided for most of the commonly used indicators of poverty and inequality.

For any complex statistic, a linearised 'indicative variable' z_i is developed such that the simple expression for its variance approximates the variance of the complex statistic. In all cases, the linearised 'indicative variable' z_i can be decomposed into two components, say: $z_i = z_{1i} + z_{2i}$, where z_{1i} is what this variable would have been if the statistic of interest were treated as a simple ratio, and z_{2i} is the extra term coming from the fact that the statistic is actually more complex.

For each poverty or inequality statistic (col 1), the table lists the two components separately (cols 2-3). The last column defines the symbols used (terms already defined in the preceding rows are generally not repeated).

Annex I: Linearised 'indicative' variables for variance estimation
Poverty, inequality and income distribution measures

Statistic	Linearised variable (λ_i)		Notes
	constant parameters	additional term	
Mean Equivalised Income $\bar{y} = \sum w_i \cdot y_i$	$(y_i - \bar{y})$	–	$\sum w_i = 1$
Functions of ratios			
Ratio $\lambda = \bar{y}/\bar{x}$	$\frac{1}{\bar{x}}.((y_i - \bar{y}) - \lambda.(x_i - \bar{x})) = \left(\frac{y_i - \lambda.x_i}{\bar{x}}\right)$	-	
Difference or sum of ratios $\lambda = r \mp r' = \frac{\bar{y}}{\bar{x}} \mp \frac{\bar{y}'}{\bar{x}'}$	$\left(\frac{y_i - r.x_i}{\bar{x}}\right) \mp .\left(\frac{y'_i - r'.x'_i}{\bar{x}'}\right) = r_i \mp r'_i$	-	
Double ratio $\lambda = r/r' = \frac{\bar{y}/\bar{x}}{\bar{y}'/\bar{x}'}$	$\frac{1}{r'}.\left(\left(\frac{y_i - r.x_i}{\bar{x}}\right) - \lambda.\left(\frac{y'_i - r'.x'_i}{\bar{x}'}\right)\right) = \frac{1}{r'}.(r_i - \lambda.r'_i)$	-	
Product of ratios $\lambda = r * r' = \frac{\bar{y}}{\bar{x}} * \frac{\bar{y}'}{\bar{x}'}$	$r'.\left(\frac{y_i - r.x_i}{\bar{x}}\right) + r.\left(\frac{y'_i - r'.x'_i}{\bar{x}'}\right) = r'.r_i + r.r'_i$	-	
Any function of ratios $\lambda = \lambda(r_1, r_2, r_{3,\ldots})$	$\lambda_i = \left(\frac{\partial\lambda}{\partial r_1}\right).r_{1i} + \left(\frac{\partial\lambda}{\partial r_2}\right).r_{2i} + \left(\frac{\partial\lambda}{\partial r_3}\right).r_{3i} + \ldots$	-	

Income Exponent $E = \sum w_i . \exp\left(- \frac{y_i}{\bar{y}} \right)$	$e^{-\frac{y_i}{\bar{y}}} - E$	$Y . \left(\frac{y_i - \bar{y}}{\bar{y}} \right)$	where $Y = \sum w_i . \left(\frac{y_i}{\bar{y}} . e^{-\frac{y_i}{\bar{y}}} \right)$
Coefficient of variation (squared). $CV^2 = \sum w_i . \left(\frac{y_i - \bar{y}}{\bar{y}} \right)^2$	$\left(C_i^2 - CV^2 \right)$	$-2 . C_i . CV^2$	where $C_i = \left(\frac{y_i - \bar{y}}{\bar{y}} \right)$
Coefficient of variation (un-squared). $CV = \sqrt{\sum w_i . \left(\frac{y_i - \bar{y}}{\bar{y}} \right)^2}$	$\left(C_i^2 - CV^2 \right) / 2 . CV$	$- C_i . CV$	
Relative mean deviation $RDM = \sum w_i . \left(\frac{\lvert y_i - \bar{y} \rvert}{2 . \bar{y}} \right)$	$\left(\frac{\lvert y_i - \bar{y} \rvert}{2 . \bar{y}} \right) - RDM$	$\left(y_i - \bar{y} \right) . \frac{1}{2 . \bar{y}^2} . \sum w_i . \left(\delta \left(y_i < \bar{y} \right) . y_i - \delta \left(y_i > \bar{y} \right) . y_i \right)$	
Theil's measure of general entropy $T = \sum w_i . \left(\frac{y_i}{\bar{y}} . \ln \left(\frac{y_i}{\bar{y}} \right) \right)$	$\frac{y_i}{\bar{y}} . \ln \left(\frac{y_i}{\bar{y}} \right) - T$	$- \left(\frac{y_i - \bar{y}}{\bar{y}} \right) . \left(1 + T \right)$	
Atkinson with $\varepsilon = 1$ $ATK = 1 - e^a = 1 - e^{\sum w_i a_i}$ where $a_i = \ln \left(\frac{y_i}{\bar{y}} \right)$	$- e^a . \left(a_i - a \right)$	$e^a . \left(\frac{y_i - \bar{y}}{\bar{y}} \right)$	

Atkinson with $\varepsilon \neq 1$ $\mathrm{ATK} = 1 - (\mathrm{a})^{\frac{1}{1-\varepsilon}} = 1 - \left(\sum \mathrm{w_i a_i}\right)^{\frac{1}{1-\varepsilon}}$ where $\mathrm{a_i} = \mathrm{w_i}.\left(\frac{\mathrm{y_i}}{\bar{\mathrm{y}}}\right)^{1-\varepsilon}$	$-\frac{\mathrm{a}^{\frac{+\varepsilon}{1-\varepsilon}}}{1-\varepsilon}.(\mathrm{a_i} - \mathrm{a})$	$\mathrm{a}^{\frac{1}{1-\varepsilon}}.\left(\frac{\mathrm{y_i} - \bar{\mathrm{y}}}{\bar{\mathrm{y}}}\right)$	
Gini $G = \sum w_i \cdot \left(2 \cdot \frac{y_i - \bar{y}}{\bar{y}} \cdot W_i\right);\ W_i = \sum_{j=1}^{i} w_j$	$2 \cdot W_i \cdot \left(\frac{y_i - \bar{y}}{\bar{y}}\right) - G$	$-(1+G) \cdot \left(\frac{y_i - \bar{y}}{\bar{y}}\right) + 2 \cdot (W_i - Y_i) - G$	$Y_i = \sum_{j=1}^{i} w_j \cdot \left(\frac{y_j}{\bar{y}}\right)$
α ***-th*** **quantile** $\mathrm{y}_\alpha \mid \sum \mathrm{w_i} \delta(\mathrm{y_i} \leq \mathrm{y}_\alpha) = \alpha$	-	$-\left(\frac{\alpha_\mathrm{i} - \alpha}{\mathrm{f}_\alpha}\right)$	where $\mathrm{dF}/\mathrm{dy}_\alpha = \mathrm{f}_\alpha$ and $\alpha_\mathrm{i} = \delta(\mathrm{y_i} \leq \mathrm{y}_\alpha)$
Quantile ratio $\lambda = \mathrm{y}_\beta / \mathrm{y}_\alpha;$	-	$-\frac{1}{\mathrm{y}_\alpha}.\left(\left(\frac{\beta_\mathrm{i} - \beta}{\mathrm{f}_\beta}\right) - \lambda.\left(\frac{\alpha_\mathrm{i} - \alpha}{\mathrm{f}_\alpha}\right)\right)$	where $\mathrm{dF}/\mathrm{dy}_\beta = \mathrm{f}_\beta$ and $\beta_\mathrm{i} = \delta(\mathrm{y_i} \leq \mathrm{y}_\beta)$
Mean equivalised income of (bottom) α ***-th*** **quantile..** $\mathrm{m}_\alpha = \sum \mathrm{w_i}.\delta(\mathrm{y_i} \leq \mathrm{y}_\alpha)\mathrm{y_i}$	$\left(\frac{\alpha_\mathrm{i}}{\alpha}\right).(\mathrm{y_i} - \mathrm{m}_\alpha)$	$-(\mathrm{y}_\alpha - \mathrm{m}_\alpha).\left(\frac{\alpha_\mathrm{i} - \alpha}{\alpha}\right)$	
Mean equivalised income of <u>top</u α ***-th*** **quantile.** $\bar{\mathrm{m}}_\alpha = \sum \mathrm{w_i}.\delta(\mathrm{y_i} > \mathrm{y}_\alpha)\mathrm{y_i}$	$\left(\frac{\alpha_\mathrm{i}}{\alpha}\right).(\mathrm{y_i} - \bar{\mathrm{m}}_\alpha)$	$-(\mathrm{y}_{\bar{\alpha}} - \bar{\mathrm{m}}_\alpha).\left(\frac{\alpha_\mathrm{i} - \alpha}{\alpha}\right)$	
Income share of (bottom) α ***-th*** **quantile (Lorenz curve ordinate)** $\mathrm{s}_\alpha = \frac{1}{\bar{\mathrm{y}}} \sum \mathrm{w_i} \delta(\mathrm{y_i} \leq \mathrm{y}_\alpha).\mathrm{y_i}\ \cdot$	$\left(\frac{y_i}{\bar{y}} \cdot \alpha_i - s_\alpha\right)$	$-s_\alpha\left(\frac{y_i - \bar{y}}{\bar{y}}\right)$ $-\left(\frac{y_\alpha}{\bar{y}}\right).(\alpha_i - \alpha)$	
Income share of top α ***-th*** **quantile.** $\bar{s}_\alpha = \frac{1}{\bar{y}} \Sigma w_i.(\delta(y_i > y_\alpha).y_i)$	$\left(\frac{y_i}{\bar{y}} \cdot \alpha_i - \bar{s}_\alpha\right).$	$-\bar{s}_\alpha\left(\frac{y_i - \bar{y}}{\bar{y}}\right)$ $-\left(\frac{y_{\bar{\alpha}}}{\bar{y}}\right).(\alpha_i - \alpha)$	

Share ratio (top:bottom α *-th* quantile) $\lambda = \frac{\bar{s}_\alpha}{s_\alpha} = \frac{\sum w_i \cdot \delta(y_i > y_{1-\alpha})}{\sum w_i \cdot \delta(y_i \le y_\alpha)}$	$\frac{1}{s_\alpha} \cdot \left(\frac{y_i}{\bar{y}}\right) \cdot (\bar{\alpha}_i - \lambda \cdot \alpha_i)$	$-\frac{1}{s_\alpha} \cdot \left[\left(\frac{y_{1-\alpha}}{\bar{y}}\right) \cdot (\bar{\alpha}_i - \alpha) - \lambda \cdot \left(\frac{y_\alpha}{\bar{y}}\right) \cdot (\alpha_i - \alpha)\right]$	$\alpha_i = \delta(y_i \le y_\alpha)$ $\bar{\alpha}_i = \delta(y_i > y_{1-\alpha})$ Example: α=0.20 → S80/S20
Income share between α *-th* and β -th quantiles . $s_{\beta\alpha} = (s_\beta - s_\alpha).$	$\left(\frac{y_i}{\bar{y}}\right).(\delta_i - s_{\beta\alpha})$	$-\left(\frac{y_\beta}{\bar{y}}\right).(\beta_i - \beta) + \left(\frac{y_\alpha}{\bar{y}}\right).(\alpha_i - \alpha)$	$\delta_i = \delta(y_\alpha < y_i \le y_\beta) = (\beta_i - \alpha_i)$
Head Count Ratio $p = \sum w_i \cdot (\delta(y_i \le y_p)) = \sum w_i \cdot p_i$ $\sum w_i \cdot \delta(y_i \le y_\alpha) = \alpha$ $y_p = \beta \cdot y_\alpha$	$(p_i - p)$	$-\beta \cdot f_p \cdot \left(\frac{\alpha_i - \alpha}{f_\alpha}\right)$	$\alpha = 0.50 \quad \beta = 0.60$ $\alpha_i = \delta(y_i \le y_\alpha)$ *f_p , f_α density functions*
Poverty rate (mean based) $p = \Sigma w_i.(\delta(y_i \le y_p))$	$(p_i - p)$	$+\beta.f_p.(y_i - \bar{y})$	Poverty line $y_p = \beta.\bar{y}$
Mean income of the poor $\bar{y}_p = \frac{\sum w_i \cdot y_i \cdot p_i}{\sum w_i \cdot p_i}$	$(y_i - \bar{y}_p).\frac{p_i}{p}$	$(y_p - \bar{y}_p).\frac{f_p}{p} \cdot \beta_i$	
Foster-Greer-Thorbecke index (FGT) $y_p = \beta y_\alpha$. $FGT = \Sigma w_i.\left(\delta(y_i \le y_p).\left(\frac{y_p - y_i}{y_p}\right)^\varepsilon\right)$ $= \Sigma w_i.(p_i.\varepsilon_i)$	$p_i.\varepsilon_i - FGT$	$-\beta.\left(\frac{\alpha_i - \alpha}{f_\alpha}\right).\frac{\partial T}{\partial y_p}$ $\varepsilon = 1 : \frac{\partial T}{\partial y_p} = \left(\frac{1}{y_p^2}\right).\Sigma w_i.p_i.y_i$ $\varepsilon = 2 : \frac{\partial T}{\partial y_p} = \left(\frac{2}{y_p^2}\right).\Sigma w_i.p_i.y_i.\left(\frac{y_p - y_i}{y_p}\right)$	$\frac{\partial T}{\partial y_p} = \Sigma w_i.\left(p_i.\frac{\partial \varepsilon_i}{\partial y_p}\right),$ $= \left(\frac{\varepsilon}{y_p^2}\right).\Sigma w_i.p_i.y_i.\left(\frac{y_p - y_i}{y_p}\right)^{\varepsilon - 1}$ $y_{p_i} = -\beta.\left(\frac{\alpha_i - \alpha}{f_\alpha}\right)$
FGT mean-based poverty line $y_p = \beta\bar{y}$ As above except for definition of y_{pi}	$p_i.\varepsilon_i - FGT$	$+\beta.(y_i - \bar{y}).\frac{\partial T}{\partial y_p}$	$y_{p_i} = \beta \cdot (y_i - \bar{y})$

FGT/H, e=1, Average poverty gap ratio(APGR) $\mathrm{APGR} = \frac{\sum \mathrm{w_i} \cdot \delta(\mathrm{y_i} \le \mathrm{y_p}) \cdot \left(\frac{\mathrm{y_p} - \mathrm{y_i}}{\mathrm{y_p}} \right)}{\sum \mathrm{w_i} \cdot \delta(\mathrm{y_i} \le \mathrm{y_p})} = \frac{\sum \mathrm{w_i} \cdot \mathrm{p_i} \cdot \varepsilon_i}{\sum \mathrm{w_i} \cdot \mathrm{p_i}}$	$\frac{p_i}{p} \cdot (\varepsilon_i - APGR)$	$y_{p_i} \cdot \left[\left(\frac{\bar{y}_p}{y_p^2} \right) - \frac{APGR}{p} \cdot f_p \right]$	$y_{p_i} = -\beta \cdot \left(\frac{\alpha_i - \alpha}{f_\alpha} \right)$
Relative median at-risk-of-poverty gap $\mathrm{E} = \left(\frac{\mathrm{y_p} - \mathrm{y_m}}{\mathrm{y_p}} \right)$ where $\mathrm{y_m} \mid \sum \mathrm{w_i} \cdot \delta(\mathrm{y_i} \le \mathrm{y_m}) = 0.5 \cdot \mathrm{p}$	–	$\frac{1}{y_p} \cdot \left[y_{\gamma_i} - \left(\frac{y_m}{y_p} \right) \cdot y_{p_i} \right]$	$\alpha = 0.50 \quad \beta = 0.60$ $y_{p_i} = -\beta \cdot \left(\frac{\alpha_i - \alpha}{f_\alpha} \right)$; $h_i = (p_i - p) + (f_p \cdot y_{p_i})$ $\gamma_i = \delta(y_i \le y_m)$; $y_{\gamma_i} = \frac{1}{f_m} \cdot [0.5 \cdot h_i - (\gamma_i - 0.5 \cdot p_i)]$
S=SEN **Median-based poverty line** $\mathrm{y_p} = \beta . \mathrm{y}_\alpha$ $S = 2.\Sigma w_i \cdot \left(\left(\frac{y_p - y_i}{y_p} \right) . (1 - G_i) . p_i \right)$;	$2.\mathrm{x_i}.(1 - \mathrm{G_i}).\mathrm{p_i} - \mathrm{S}$	$-\beta . \left(\frac{\mathrm{p} - \mathrm{S}}{\mathrm{y_p}} \right) . \left(\frac{\alpha_\mathrm{i} - \alpha}{\mathrm{f}_\alpha} \right)$ $+ 2.\mathrm{p_i}.(\mathrm{G_i} - \mathrm{Y_i}) - (\mathrm{S} - \mathrm{X})$	$G_i = \frac{1}{p} . \Sigma^i_{j=1} w_j . p_j + o(1/p.n).$ $Y_i = \sum_{j=1}^{i} w_j \cdot p_j \cdot \left(\frac{y_j}{y_p} \right)$ $\mathrm{x_i} = \left(\frac{\mathrm{y_p} - \mathrm{y_i}}{\mathrm{y_p}} \right).$ $\mathrm{X} = 2.\mathrm{p}.\left(\frac{\mathrm{y_p} - \bar{\mathrm{y}}_\mathrm{p}}{\mathrm{y_p}} \right), \quad \bar{\mathrm{y}}_\mathrm{p} = \frac{\Sigma \mathrm{w_i}.\mathrm{p_i}.\mathrm{y_i}}{\Sigma \mathrm{w_i}.\mathrm{p_i}}$
SEN **Mean-based poverty line** $\mathrm{y_p} = \beta . \bar{\mathrm{y}}$ As above except for definition of y_{pi}	$2.\mathrm{x_i}.(1 - \mathrm{G_i}).\mathrm{p_i} - \mathrm{S}$	$+\beta . \left(\frac{\mathrm{p} - \mathrm{S}}{\mathrm{y_p}} \right) . (\mathrm{y_i} - \bar{\mathrm{y}})$ $+ 2.\mathrm{p_i}.(\mathrm{G_i} - \mathrm{Y_i}) - (\mathrm{S} - \mathrm{X})$	$y_{p_i} = \beta \cdot (y_i - \bar{y})$

P, Polarisation index $P = \Sigma w_i.\left(\frac{y_i}{y_m}.((1-G)-2.\delta(y_i \le y_m))\right)$ $= \frac{1}{y_m}.(\bar{y}.(1-G)-\bar{y}_{<m})$ y_m = median income; $\bar{y}_{<m}$ = mean of incomes below median G=Gini coefficient	$P_i = \frac{\partial P}{\partial y_m}.(y_m)_i + \frac{\partial P}{\partial \bar{y}}.\bar{y}_i + \frac{\partial P}{\partial \bar{y}_{<m}}.(\bar{y}_{<m})_i + \frac{\partial P}{\partial G}.G_i$ $\frac{\partial P}{\partial y_m} = -\frac{1}{y_m^2}.(\bar{y}.(1-G)-\bar{y}_{<m}) = -\frac{P}{y_m};\ \frac{\partial P}{\partial \bar{y}} = \frac{1-G}{y_m};\ \frac{\partial P}{\partial \bar{y}_{<m}} = -\frac{1}{y_m};\ \frac{\partial P}{\partial G} = -\frac{\bar{y}}{y_m}$ $(y_m)_i = -\left(\frac{m_i - m}{f_m}\right)$ with $m_i = \delta(y_i \le y_m); m = 0.5$ $\bar{y}_i = (y_i - \bar{y})$ $(\bar{y}_{<m})_i = \left(\frac{m_i}{m}\right).(y_i - \bar{y}_{<m}) - (y_m - \bar{y}_{<m}).\left(\frac{m_i - m}{m}\right)$ $G_i = 2.W_i.\left(\frac{y_i - \bar{y}}{\bar{y}}\right) - G - (1+G).\left(\frac{y_i - \bar{y}}{\bar{y}}\right) + 2.(W_i - Y_i) - G$		
CHU, Clark-Hamming-Ulph $\lambda = \sum w_i \cdot \left[1 - \left(\frac{y_i}{y_p}\right)^{\beta}\right] \cdot p_i$ 0<β<1	$\left\{\left[1 - \left(\frac{y_i}{y_p}\right)^{\beta}\right] \cdot p_i - \lambda\right\}$	$\frac{\beta}{y_p} \cdot (p - \lambda) \cdot y_{p_i}$	$y_{p_i} = -\beta \cdot \left(\frac{\alpha_i - \alpha}{f_\alpha}\right)$
Watts $\lambda = \sum w_i \cdot \left[-\ln\left(\frac{y_i}{y_p}\right)\right] \cdot p_i$	$-p_i \cdot \ln\left(\frac{y_i}{y_p}\right) - \lambda$	$\frac{p}{y_p} \cdot y_{p_i}$	
CDS, Constant distribution sensitivity $\lambda = \sum w_i \cdot \{\exp[\beta \cdot (y_p - y_i)] - 1\} \cdot p_i$ β>0	$\{\exp[\beta \cdot (y_p - y_i)] - 1\} \cdot p_i - \lambda$	$\beta \cdot (\lambda + p) \cdot y_{p_i}$	

References

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A., Patak, Z. (1994). Use of estimation functions for interval estimation from complex surveys. *Journal of American Statistical Association*, 89, 1035-1043.

Binder, D.A., Kovacevic, M.S. (1995). Estimating some measures of income inequality from survey data: an application of the estimation equation approach. *Survey Methodology*, 21, 137-145.

Brick, J.M., Morganstein, D. (1997). Computing sampling errors from clustered unequally weighted data using replication: WesVarPC. Bulletin of the International Statistical Institute, Book 1, 479-482.

Demnati, A., Rao, J.N.K. (2004). Linearised variance estimators for survey data. *Survey Methodology*, 30(1), 17-26.

Deming, W.E. (1943). *Statistical Adjustment of Data*. New York: Wiley.

Deville, J.C. (1999). Variance estimation for complex statistics and estimators: linearisation and residual techniques. *Survey Methodology*, 25(2), 193-203.

Deville, J.C., Sarndal C.E. (1994). Variance estimation for the regression imputed Horwitz – Thompson Estimator. *Journal of Official Statistics*, 10, 381 – 394.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46, 477-480.

Efron, B., Stein, C. (1981). The Jackknife estimate of variance. *Annals of Statistics*, 9, 586-596.

Fay, B. E. (1994). Analyzing imputed survey datasets with model assisted estimators. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 900 – 905.

Haziza, D., Rao, J.N.K. (2003). Inference for population means under unweighted imputation for missing survey data. *Survey Methodology*, 29(1), 81-90

Kendall, M.G., Stuart, A. (1958). *The Advanced Theory of Statistics*, Vol I. London: Charles Griffin.

Keyfitz, N. (1957). Estimation of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association,* 52, 503-510.

Kish, L., Frankel, M. (1974). Inferences from complex samples. *Journal of the Royal Statistical Society*, B/36, 1-37.

Kovacevic, M.S., Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization – the estimation equation approach. *Journal of Official Statistics*, 13(1), 41-58.

Kovacevic, M.S., Yung, W. (1997). Variance estimation for measures of income inequality and polarization – an empirical study. *Survey Methodology*, 23(1), 41-52.

Lee, H., Rancourt, E., Sarndal, C.E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231 – 243.

Lee, H., Rancourt, E., Sarndal, C.E. (1995). Variance estimation in the presence of imputed data for the generalized estimation system. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 384 – 389.

Preston, I. (1995). Sampling distribution of relative poverty statistics. *Applied Statistics*, 44, 91-99.

Rancourt, E., Sarndal, C.E., Lee, H. (1994). Estimation of the variance in the presence of nearest - neighbor imputation. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 888 – 893.

Rao, J.N.K. (1979). On deriving mean squared errors and their non-negative unbiased estimators in finite population sampling. *Journal of the Indian Statistical Association*, 17, 125-136.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434), 499 – 50.

Rao, J.N.K., Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79, 4, 811-822.

Rao, J.N.K., Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika,* 82(2), 453 – 460.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys,* New York: Wiley.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1(4), 381-397.

Shao, J., Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association,* 94(445), 254-26.

Sitter, R.R., Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *The Canadian Journal of Statistics,* 25(1), 61-73

Tukey, J. (1958). Bias and confidence in not-quite-large samples. *Annals of Mathematical Statistics*, 29, 614.

Verma, V. (1993). *Sampling Errors in Household Surveys: A Technical Study*. New York: United Nations Department for Economic and Social Information and Policy Analysis, Statistical Division, INT-92-P80-15E.

Verma, V., Lê, T. (1996). An Analysis of Sampling Errors for the Demographic and Health Surveys. *International Statistical Review*, 64(3), 265-294.

Verma, V., Scott, C., O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, A/143(4), 431-473.

Woodruff, R. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association,* 66(334), 411-414.

Yung, W., Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95(451), 903- 915.

Zheng, B. (2001). Statistical inference for poverty measures with relative poverty lines. *Journal of Econometrics*, 101, 337-356.