INT-92-P80-15E

UNITED NATIONS

DEPARTMENT FOR ECONOMIC AND SOCIAL INFORMATION AND POLICY ANALYSIS

STATISTICAL DIVISION
NATIONAL HOUSEHOLD SURVEY CAPABILITY PROGRAMME

SAMPLING ERRORS IN HOUSEHOLD SURVEYS

UNFPA/UN/INT-92-P80-15E

NATIONAL HOUSEHOLD SURVEY CAPABILITY PROGRAMME

SAMPLING ERRORS IN HOUSEHOLD SURVEYS

UNITED NATIONS
DEPARTMENT FOR ECONOMIC
AND SOCIAL INFORMATION
AND POLICY ANALYSIS

STATISTICAL DIVISION

NEW YORK, 1993.

PREFACE

This is one of a series of technical studies produced under the auspices of the National Household Survey Capability Programme of the United Nations Statistical Division (UNSTAT) of the Department for Economic and Social Information and Policy Analysis. The series is designed to assist countries, particularly developing countries, in planning and implementing household surveys. A number of studies have been published to provide reviews of issues and procedures in specific areas of household survey methodology and in selected subject areas.

Irrespective of their particular source, all statistical data are subject to errors of various types. As a component of the total error, sampling error is a measure of the uncertainty in the results arising from the fact that inferences about the whole population are drawn from observations confined to a sample. Information on sampling errors is needed for the correct interpretation of sample survey results and for improving survey design.

This study aims to provide a basic understanding of the practical procedures for computing sampling errors and guidelines on how to analyse and utilise this information. Topics are discussed with numerous illustrations to ensure that the study is of maximum benefit to sampling practitioners.

The study was prepared by Mr. Vijay Verma who assisted the United Nations as a consultant. It was edited by Mr. Edmundo Berumen-Torres. Among the reviewers, special appreciation is extended to Professor Leslie Kish and Mr. Christopher Scott for valuable suggestions leading to improvement of this work. The study was prepared with financial support from the United Nations Population Fund.

SAMPLING ERRORS IN HOUSEHOLD SURVEYS

CONTENTS

PART II - ANALYSIS

CHAPTER 5 DECOMPOSITION OF THE TOTAL VARIANCE

CHAPTER 6 DATA REDUCTION AND MODELLING

1

INTRODUCTION

1.1 THE CONTEXT

It is widely recognised as good practice for survey reports to include detailed information on the sampling variability of survey estimates. Information on sampling errors is needed both for the correct interpretation of survey results and for evaluating and improving survey design. Yet no such information is included in many reports based on sample surveys. The primary reason for this state of affairs is a lack of appreciation among many producers as well as users of statistics of the significance of information on sampling errors.

The present study is one of a series of Technical Studies produced under the auspices of the National Household Survey Capability Programme with the objective of improving the quality of household survey work. It aims to provide a basic understanding of the practical procedures of computing sampling errors, and guidelines on how to analyse and utilise this information in the context of large-scale household surveys. Various topics are discussed in sufficient technical detail with numerous illustrations, and in a reasonably self-contained manner, to ensure that the study is of maximum benefit to sampling practitioners in developing countries.

While we have tried to make this study as self-contained and technically comprehensive as possible, it has been assumed that the reader is familiar with the basic concepts of sampling theory and survey practice. Of course, concepts and procedures which directly pertain to the discussion at hand are clearly defined and explained to the extent possible, and a brief introduction to some basic ideas is given in this chapter. In the main, however, reference must be made to the much more comprehensive explanations available in many standard texts on sampling methods and in the literature generally.

To compute sampling errors for diverse survey estimates, it is necessary to have computer software specifically designed for the purpose. In Chapter 5 a brief review is given of some of the programs available at the time the present report was prepared.

1.2 OUTLINE OF THE CONTENT

Part I of this study provides a technical description of the various practical procedures for computing sampling errors in large-scale surveys with complex designs. Chapter 2 describes, with numerical examples, the most commonly used method which is based on comparisons between primary selections within each stratum of a multi-stage stratified design. Chapter 3 describes various methods based on the idea of sample replications, which are more readily extended to complex statistics. Diverse practical issues in the implementation of these variance estimation procedures are addressed in Chapter 4, including a brief review of the available software for the purpose.

Part II considers the analysis and use of the information on sampling errors. Chapter 5 is concerned with decomposition of the overall sampling error into components which are valuable in analysis and better utilization of the information. This includes as its basis the decomposition into (i) the part of the overall sampling error which would be obtained in a SRS, and (ii) the design effect which measures the effect on sampling error of various complexities of the sample design. Each part can be further decomposed, for instance to isolate the effect of sample size and weighting. Chapter 6 considers, more comprehensively and with many illustrations, the 'modelling' of sampling errors to develop measures which are 'portable' for use from one statistic and situation to another. On this basis, sampling errors for diverse subclasses and differences between classes can be related to those for the total sample. Chapter 7 provides, again with many illustrations, issues relating to the presentation of sampling errors to suit the requirements of different categories of users.

1.3 SOME BASIC CONCEPTS AND PROCEDURES

This section briefly reviews some basic aspects of survey structure and design to which repeated reference will be made in the course of discussion on sampling errors in the following chapters.

Multi-purpose Surveys; Complex Design and Estimation

The requirements and procedures for computing sampling errors have to be determined on the basis of the type of application required.

A common feature of national household surveys is their multipurpose nature. A typical survey is multipurpose in several respects. It may involve many types of substantive variables; for any variable, different types of statistics such as estimates of aggregates or totals, proportions or percentages, means, rates and ratios, differences and other functions of ratios may be involved. Statistics may be required not only at the national level, but also separately for various geographical domains such as urban and rural areas and regions, and for numerous other subclasses or groups in the population. The need for comparisons among groups in the population can vastly increase the number of separate statistics involved; in analytical surveys other, more complex measures of relationship such as regression and correlation coefficients may also be encountered. Another dimension of variation is the different types of units of analysis in the survey, such as individual persons, households, or communities. Some household surveys also involve analysis in terms of non-household units such as agricultural holdings and household enterprises, or various subunits within households such as earning, spending and family units.

A typical national household survey may involve a sample of several thousand households selected from area units in a number of stages with stratification at each stage. Special selection procedures such as systematic sampling and selection with probability proportional to size (PPS) may be involved. The sampling rates may differ between different domains such as urban and rural areas or regions of the country. Multi-phase sampling and overlaps or rotation between samples are other examples of complexity. The sample data may be weighted. More complex composite, synthetic, seasonally adjusted estimates, etc, may also be involved, though the use of complex estimation procedures tends to be less common in developing countries, at least in part due to the lack of auxiliary information required for their application.

All these aspects of complexity and diversity tend to be even more pronounced in the case of programmes of household surveys, where individual surveys may have different objectives and content, different designs and structures, and diverse operational and substantive linkages.

Multi-stage Sampling

In household surveys in developing countries (and in many developed countries as well), samples of households and persons are usually selected in a number of sampling stages. For instance, the whole country may be divided into area units such as localities or census enumeration areas (EAs), and a sample of these areas selected at the first stage. The type of units selected at the first stage are called primary sampling units (PSUs). For the first stage of selection, a frame of PSUs is needed which lists the units covering the entire country exhaustively and without overlaps, and which also provides information for the selection of units efficiently. Such a frame is called the primary sampling frame (PSF). The next (second) stage may consist of dividing each of the PSUs selected at the first stage into smaller areas such as blocks, and selecting one or more of these second stage units (SSUs) from each selected PSU. This process may continue till a sample of sufficiently small ultimate area units (UAUs) is obtained. Finally, in each selected UAU, individual households may be listed and a sample selected with households as the ultimate sampling units (USUs). In the survey, information may be collected and analyzed for the USUs themselves; or for other types of units ('elements') associated with the selected USUs, such as individual persons within sample households.

In a multi-stage sample, the probability of selection of an ultimate unit is the product of probabilities at the various stages of selection. It is possible (and common) to have varying probabilities at different stages, but balanced such that the overall probability of all ultimate units is uniform. A common procedure is to select the PSUs (and other higher stage units) with probabilities proportional to some measure of their size (PPS), and to obtain an equal probability sample of elements ('epsem') by selecting the USUs with probabilities which compensate for differences at the preceding stages.

Stratification

Stratification means dividing the units in the population into groups and then selecting a sample independently within each group. This permits separate control over the design and selection of the sample within each stratum, such as urban-rural areas or regions of a country. This means that different parts of the population can be sampled differently, using different sampling rates and designs as necessary. The separation may also be retained at the stage of sample implementation and estimation and analysis, but this is not essential to the idea of stratification. It is common for instance to pool the results from different strata to produce estimates for the whole population, or for major parts or 'domains' of the population each of which is composed of a number of strata. However, it is important to note in the present context that in estimating variances, it is necessary to take fully into account the stratification as it affects the magnitude of the sampling error.

Probability Sampling; Measurable Samples

It is necessary, at least in the context of 'official' statistics, that surveys are based on probability and measurable samples. A probability sample means that every element in the population is given a known and non-zero chance of being selected into the sample. To obtain a probability sample, certain proper procedures must be followed at the selection and implementation stages. The sample has to be selected from a frame, representing all elements in the population, by a suitable randomisation process which gives each unit the specified probability of selection; in addition, in estimating population values from the sample, the data from each unit in the sample should be weighted in accordance with the unit's chance of selection. The major strength of probability sampling is that the probability selection mechanism permits the application of statistical theory to produce valid estimates of the population values of interest, and furthermore, to examine the properties such as variance of these estimates.

A related but more demanding concept is that of measurability. A sample is called measurable if from the variability observed between units within the sample, usable estimates of the sampling variance (ie. of the variability between different possible samples) can be obtained. To be measurable, it is highly desirable that the sample be a probability sample; it should also meet certain other requirements to ensure that the sampling variability can be estimated from the observed variability between units in the one sample that is available.

Simple Random Sampling

Though usually of limited relevance in the context of household surveys in developing countries, a simple random sample (SRS) provides the point of reference against which the statistical efficiency, as well as more generally the cost and other aspects of the quality of the actual complex designs used can be evaluated. We will make constant reference to such comparison in the form of the 'design effect'. A SRS is obtained by a series of random selections applied directly to the population elements, which ensures that the chance of selection is the same not only for the elements individually, but also in all combinations of any given size.

Systematic Sampling

A common method of sample selection is to select the units systematically from a list ordered in some way. The basic idea is as follows. Suppose that an equal probability sample of n units (listings) is required from a population of N. From the list of units, preferably arranged in some useful way, one unit is selected from every $I=N/n$ units in the list. A random number r between 1 and I identifies a sequence number of the first unit selected. Then starting with r, every Ith unit may be selected, ie the sequence numbers selected being r, $r+I$, $r+2I$,....., $r+(n-1)I$. To the extent that the units in the original list appear in a random order, the resulting sample is equivalent to a random sample of the units concerned. However, existing lists are practically never randomly ordered; in any case the objective of systematic sampling is to make use of the order available to achieve a better spread of the sample according to some meaningful criterion, such as geographical location of the units. In this manner, systematic sampling provides implicit stratification; it can be regarded as stratification of the population into zones of size I, and the selection of one unit per zone or 'implicit stratum'. The widespread use of systematic sampling is also due to the great convenience of the method in many situations.

1.4 ESTIMATION

Some remarks on the basic estimation procedures used in surveys will be useful as a background to the discussion on variance estimation.

Estimating Proportions, Means or Other Ratios from a Multi-stage Sample

The most common type of estimator encountered in surveys takes the form of a ratio of two sample aggregates, say y and x:

$$\begin{aligned} y &= \sum_i y_i = \sum_i \sum_j w_{ij} \cdot y_{ij} \\ x &= \sum_i x_i = \sum_i \sum_j w_{ij} \cdot x_{ij} \\ r &= y/x \end{aligned} \tag{1.1}$$

Both the numerator (y) and denominator (x) may be substantive variables - as for example in the estimation of income per capita from a household survey, where y is the total income and x the total number of persons estimated from the survey. For ultimate unit j (a household) in PSU i, y_{ij} refers to its income and x_{ij} to its size (=number of persons, in this example). Quantity w_{ij} is the weight associated with the unit. Sophisticated methods may be used to compute the weights to be applied, but at the most basic and important level, the weights - called the design weights - are inversely proportional to the units' probabilities of selection into the sample. With the same probability given to all units, the design weights are also uniform and the sample is termed self-weighting.

Ordinary means, percentages and proportions are just special cases of the ratio estimator. In a mean, the denominator is a count variable, that is, x_{ij} is identically equal to 1 for all elements in the sample. This gives

$$\bar{y} = \frac{\sum_i \sum_j w_{ij} \cdot y_{ij}}{\sum_i \sum_j w_{ij}} \tag{1.2}$$

For a proportion (or percentage) the additional condition is that y_{ij} is a dichotomy equal to 1 or 0 depending on whether or not unit j possesses the characteristic whose proportion is being estimated.

The survey may also involve more complex statistics such as differences, weighted sums, ratios or other functions of ratios. These can be estimated in an analogous way.

Estimating Totals

The simple unbiased estimator (equation 1.3 below) is usually not satisfactory in practice for estimating population aggregates. This is especially the case for surveys with a multi-stage design and small sample size. This is because with multi-stage sampling, the resulting sample size varies at random, and therefore aggregates directly estimated from the survey can have a large sampling error. The problem is even more serious when estimates are required for population subclasses the selection of which is not explicitly controlled in the multi-stage design.

An equally important problem arises from the fact that estimates of aggregates are directly biased in proportion to the magnitude of the coverage and related errors. By contrast, this effect on estimates of ratios can be much less marked.

The appropriate procedure for estimating population aggregates is generally as follows. In place of a simple inflation of the form

$$\hat{Y} = F.y \tag{1.3}$$

the required aggregate may be expressed in the form of a ratio-type estimator

$$\hat{Y}_r = \frac{y}{x}.X \tag{1.4}$$

where y and x are estimated totals from the sample; y being the variable of interest, and x an auxiliary variable for which a more reliable population aggregate value X is available from some external source. The value and applicability of this procedure depends on several factors. Firstly, the correlation coefficient between y and x must be positive and preferably large, say greater than 0.6 or 0.7 at least. Secondly, X should be available with higher precision than the simple estimate x of the population aggregate which can be directly produced from the sample itself. Thirdly, X in the population and x in the sample should be based on essentially similar coverage and measurement; a difference between the two would introduce a bias into the estimate. This often requires that values of the variable x for individual units - unless they are simply a count of the cases, as in the case of an ordinary mean - are taken from the external source rather than directly from the measurements in the survey, though of course that must be for the actual units included in the sample.

The precondition for the use of this procedure of course is the availability of an appropriate external total X. In many situations, such a total is obtained from sources such as censuses, administrative records or very large samples, which may be considered practically free of sampling error. In such cases, the variance of (1.4) can be obtained from that of the ratio y/x (multiplied by X^2).

For this reason, and the fact that means and proportions are merely special cases, the focus in the discussion to follow is primarily on the estimation of variances of ratios (and related statistics such as differences of ratios).

<u>Weighting of the sample data</u>

In producing survey estimates, weighting of the sample data may be introduced for several reasons. The primary factor is the weighting of sample elements in inverse proportion to their selection probabilities. Additional weights are often also introduced for other reasons, such as to take into account under-coverage, non-response, and other factors resulting in departures between the sample results and the corresponding information about the population available more reliably from other sources. The issues involved in the weighting of complex surveys are themselves complex, and need not concern us here. The important point is that, however determined, the weights used in producing estimates from the survey are also relevant in estimating their variances. For this, it is essential that all information on weights be documented and preserved, preferably as an integral part of the survey data files.

<u>First and second order statistics</u>

As noted above, in a multi-stage sample the probability of selection of an ultimate unit is the product of probabilities at the various stages of selection. In estimating proportions, means and other types of ratios, it is only the ultimate sampling probabilities and not the details at various stages which matter. In fact, apart from the weights, no other complexities of the sample selection appear in this estimation. For this reason, statistics such as proportions, means and ratios are called <u>first order statistics</u>. These are distinguished from <u>second order statistics</u>, the estimation procedures for which must take into account the complexity of the sample design. Sampling variance is the prime example of the latter type of statistics. The practical implication here is that to estimate sampling errors, it is essential to have information on the structure of the sample - both on the procedures of selection and of estimation.

1.5 SAMPLING VARIANCE

The particular units which happen to be selected into a particular sample depends on chance, the possible outcomes being determined by the procedures specified in the sampling design. This means that, even if the required information on every selected unit is obtained entirely without error, the results from the sample are subject to a degree of uncertainty due to these chance factors affecting the selection of units. Sampling variance is a measure of this uncertainty.

The distribution of estimates from all possible samples with a given design (ie selection and estimation procedure) is called the <u>sampling distribution</u> of the estimator. The average of the sampling distribution, ie of all possible sample estimates weighted according to their probabilities, is called the expected value. Symbolically we may express this as follows. If p_s is the probability and y_s the estimate from a given sample s, the expected value of the estimator y is:

$$E(y) = \sum_s p_s . y_s \tag{1.5}$$

where the sum is taken over all possible samples. In many designs, p_s is a constant; for example in a simple random sample (without replacement):

$$p_s = \frac{(N-n)!.n!}{N!} \tag{1.6}$$

since the inverse of this is the total number of possible samples, each equally likely.

The variance of y is defined as:

$$Var(y) = \sum_s p_s . [y_s - E(y)]^2 \tag{1.7}$$

For various reasons, the expected or average value from all possible samples may not equal the actual population value (Y). In the absence of measurement errors, this may arise from the particular estimation procedure, in which case it is called the technical or estimation bias:

$$Bias = E(y) - Y \tag{1.8}$$

The combined effect of variance and bias is the mean square error, which is defined in terms of the squared differences of sample estimates y_s from the actual population value Y:

$$MSE(y) = \sum_s p_s . [y_s - Y]^2 = Var(y) + (Bias)^2 \tag{1.9}$$

In most well designed and implemented samples, with appropriate estimation procedures, the estimation bias is trivial.

An important result of sampling theory is that, under certain conditions, the sampling error (variability between different samples) can be estimated from the observed variability between units in the one sample that is available.

Inference from sample surveys are made in terms of probability intervals, usually confidence intervals. These intervals are defined on the basis of an assumed form of the sampling distribution, usually taken as a normal distribution. An estimated confidence interval is a range of values which contains the population value of interest with a given level of confidence (such as 68%, 95% or 99%). (Illustration 7A.(1) provides some further remarks on this important concept.)

1.6 THE IMPORTANCE OF INFORMATION ON SAMPLING ERRORS

All statistical data, irrespective of their source and method of collection, are subject to errors of various types. It is essential that results from censuses, surveys and other sources are accompanied by a description of their quality and limitations. Information on data quality is required (i) for proper use and interpretation of the data, and (ii) for evaluation and improvement of statistical design and procedures. Continued monitoring and improvement in quality of the data generated are particularly important in the case of major undertakings such as national programmes of household surveys, because such programmes are designed to generate data of great variety and complexity, and constitute the only available source of information on many topics. It is only on the basis of detailed classification by source and type that the variety of errors limiting data accuracy can be assessed and controlled. For a comprehensive review of survey errors, the reader may consult United Nations (1982).

While survey data are subject to errors from diverse sources, information on sampling errors is of crucial importance in the proper interpretation of the survey results, and in the rational design of sample surveys. Of course, sampling error is only one component of the various types of errors in survey estimates, and not always the most important component. By the same token, it is the lower (and more easily estimated) bound of the total error. A survey will be useless if this component alone becomes too large for the survey results to add useful information with any measure of confidence to what is already known prior to the survey. Furthermore, survey estimates are typically required not only for the whole population but also separately for many subgroups in the population. Generally the relative magnitude of sampling error vis a vis other types of errors increases as we move from estimates for the total population to estimates for individual subgroups and comparison between subgroups. Information on the magnitude of sampling errors is therefore essential in deciding the degree of detail with which the survey data may be meaningfully tabulated and analyzed.

Sampling error information is also essential for sample design and evaluation. For a given survey estimate, the magnitude of its sampling error depends, among other factors, on sample size and on the sample design adopted, in particular the extent to which units in the sample are clustered together and are homogeneous within clusters. To reduce sampling error, it is necessary to increase sample size and/or to reduce the degree of clustering by scattering the sample over more areas and over larger distances. At the same time, these very factors would increase survey costs, and may also increase non-sampling biases due to the greater difficulties in quality control and supervision resulting from the increased size of the operation. A balance is therefore required to minimise the total error within given resources.

Statistical efficiency is just one of the factors involved - although one which cannot be ignored. While practical constraints define, however narrowly, the class of feasible designs, choices have to be made within those on the basis of efficiency in terms of costs and variances. Some of the obvious questions to be considered relate to sample size, allocation, clustering and stratification. For example:

- Was (is) the sample size appropriate? Did the presence of large sampling errors preclude important survey objectives being met? Or alternatively, could a smaller sample have met these objectives better, perhaps by permitting a greater control of non-sampling errors?
- Was the sample allocated appropriately between different reporting domains? Was the minimum sample allocated to any domain large enough to meet the survey objectives? How did any disproportionate allocation affect the efficiency of the overall design?

- Was the degree of clustering of sample units too high, or too low, on the basis of its effect on costs, variances and control of non-sampling errors? How much cost and trouble were saved by introducing additional sampling stages, and what was their contribution to the total sampling error?
- In terms of their sampling error, what were the most critical variables in determining sample size and design?

Generally the practical constraints are not rigorously binding in the sense of completely determining the sample design; data relating to sampling errors and costs provide, at least in principle, the decisive evidence on important aspects of design such as those noted above. Furthermore, even in the absence of data on costs, considerable progress can be made by looking at sampling errors alone.

1.7 PRACTICAL METHODS FOR VARIANCE ESTIMATION

Given the complexity of the designs, and diversity and volume of statistics encountered in national household surveys, procedures for computing sampling errors, to be practicable, must meet some basic requirements:

[1] First of all, the variance estimation procedure must take into account the actual, complex structure of the design and estimation procedures since these aspects can greatly affect the magnitude of the sampling errors involved.

[2] At the same time, however, the procedures should be general and flexible to be applicable to diverse designs. This is particularly important in the case of national programmes of household surveys where individual surveys may differ in design and procedures.

[3] The procedure should be convenient and economical for large scale application: for producing results for diverse variables, type of statistics and subclasses in large, complex surveys.

[4] Generally, any computation procedure requires some basic assumptions about the nature of the sample design for the procedure to be applicable. Preferably these requirements should not be too restrictive. Even so, designs which have to be adopted in practice hardly ever meet these requirements exactly. It is desirable therefore that the method adopted is reasonably <u>robust</u> against departures of the actual design from the 'model' assumed in the computational method.

[5] The method should be economical in terms of the effort and cost involved, including technical as well as computer resources.

[6] The procedure should have desirable statistical properties, such as small variance of the variance estimates generated, small bias and/or mean square error, and accuracy in the probability levels with which the estimated confidence levels actually cover the population parameters of interest. These statistical requirements, however, need to be qualified in relation to the practical requirements of economy, generality and flexibility noted above. The objective is not to seek theoretical perfection, but practical methods with acceptable accuracy in relation to the uses made of the information on sampling errors. Information on sampling errors can be useful in practice even if the degree of precision with

which errors for individual statistics are estimated is not high. The same criterion of the priority of practicality over theoretical exactness should determine the choice among different methods.

[7] Finally, a most basic consideration in the choice of a method is the availability of suitable computer software for its application. While larger and more developed statistical organizations are able to develop and maintain their own software for meeting specific needs, organizations with fewer resources often have to rely on general purpose software developed elsewhere. The availability and reliable maintenance of such software can be a much more important consideration in the choice of a particular approach to variance computation than moderate differences in the cost or theoretical properties of the methods involved.

An additional desirable, though more difficult and not always feasible, objective concerns the decomposition of the overall sampling error into its various components associated with different stages of sample selection and other aspects of the design and estimation procedure. Such decomposition can be valuable in sample design. However, priority generally has to be given to the economical production of overall magnitudes of the sampling error for diverse variables and subclasses, over its analysis into components.

To meet the above requirement the main approach has been to develop general methods applicable to most statistics and the diversity of designs encountered in large-scale sample surveys.

Clearly, the choice of the actual procedure has to be determined by the type of application required. In contrast to the situation sketched above for household surveys in developing countries, we may also note for instance that, at least in more developed countries with high percentage of households with telephones, smaller and simpler samples are becoming common. In a sense, there is a shift in complexity away from the sample structure to more sophisticated estimation procedures making greater use of auxiliary or external information to improve results of the sample survey. This difference in emphasis can have a bearing on the approach to sampling error computation: the procedures adopted may need to take into account more carefully, and to the extent possible quantify separately, the effect of various steps in the estimation procedure on precision of the results.

PART I
COMPUTATION

2

COMPUTING SAMPLING ERRORS: COMPARISON AMONG PRIMARY SELECTIONS

2.1 INTRODUCTION

The basis of practical procedures: Replicated Variance Estimation

The theory of 'independent replicated variance estimators' (Mahalanobis, 1944) provides the basis for most practical approaches to variance estimation, though in application to complex situations, additional assumptions and approximations are involved. The basic theory may be stated as follows. Suppose that y_j are a set of random uncorrelated variables with a common expectation $\bar{Y}$. Then the mean $\bar{y}$ of n values y_j

$$\bar{y} = \sum_j y_j / n \tag{2.1}$$

has an expected value equal to $\bar{Y}$, and its variance is given by:

$$var(\bar{y}) = s^2/n; \quad where\ s^2 = \sum_j (y_j - \bar{y})^2/(n-1) \tag{2.2}$$

The most obvious example of the above is a simple random sample (SRS) of elements selected with replacement, where y_j represent values of a certain variable for individual elements j. The same idea can be applied to the more general situation when "j" refers not to individual elements but to any set of elements uncorrelated to others in the sample, and "y_j" to any complex statistic defined for each set j. The requirement is that the y_j are uncorrelated and have a common expectation. In practice this means that the sets should be selected and observed independently, following the same selection, measurement and estimation procedures.

While the straightforward approach can be applied more or less as sketched above in certain situations (for example in a sample with many units selected independently, at random), its application to practical designs generally requires additional assumptions and approximations, resulting in various types of procedures. Essentially, variance estimation requires partitioning a given sample to produce several comparable estimates of the same population parameter, the variability among which provides a valid measure of the sampling error. Basic approaches to variance estimation can be distinguished in terms of the manner in which the sample partitions to be compared are created.

Drawing on this basic idea, two broad practical approaches to the computation of sampling errors may be identified:

[1] Computation from comparisons among certain aggregates for primary selections within each stratum of the sample.

[2] Computation from comparisons among estimates for replications of the sample, each of which reflects the structure of the full sample, including its stratification.

This chapter is concerned with approach [1]. Various forms of application of the second approach will be discussed in Chapter 3*.

Method [1] involves comparison among independent primary selections comprising the sample. The term primary selection (PS) refers to the aggregate of elements selected within any primary sampling unit (PSU) in the sample. (Other terms such as 'ultimate cluster' or 'replicate' have also been used for the same thing.) The basic model of the method is that the sample is divided into independent partitions or strata, from each of which two or more independent primary selections are made such that each selection or replicate provides a valid estimate of the stratum total or a similar linear statistic. By linear statistic is meant a statistic which can be simply aggregated - with appropriate weights as necessary - to the PSU level, then to the stratum level, and finally across strata to produce the overall estimate or total for the survey population. Statistics such as ratios of two totals computed at lower levels cannot be aggregated to higher levels in this way; hence they are termed non-linear statistics.

In the usual way, the mean square deviation between the independent estimates from the primary selections provides a measure of variance within the stratum. The stratum estimates and variances can then be aggregated to produce the corresponding quantities for the whole population. With independent primary selections, the details of the sample design within PSUs do not complicate the variance estimation formulae, imparting great simplicity and generality to this technique. Directly the method is applicable to linear statistics only; generalization to more complex (non-linear) statistics requires linearisation approximations through which the complex statistics can be expressed as linear functions of simple (linear) aggregates with constant coefficients. In view of the general application to ratios and other complex statistics encountered in most surveys, the method is referred to as the linearisation method.

For good summaries of the various practical procedures for variance estimation, see Kalton (1977) and Rust (1985).

*Method [1] is known by various names such as 'linearisation', 'Taylor series', 'ultimate clusters', or 'primary selection comparison' (PSC) method. The reasons behind some of these names will become clearer later in this chapter. While the last mentioned name; PSC, is most appropriate in describing the basic approach, we will generally refer to the method as 'linearisation' in view of common usage of that term in the literature.

2.2 DESCRIPTION OF THE METHOD

Estimating Variance of a Linear Statistic

The basic assumptions about the sample design are that

[1] The sample selection is independent between strata.

[2] Two or more primary selections are drawn from each stratum.

[3] These primary selections are drawn at random, independently and with replacement. This last condition requires sampling with replacement at all stages, but can be partly relaxed as noted later.

Given independent sampling with-replacement of two or more PS's per stratum, the simple replicated sampling theory can be used to estimate the variance of linear statistics, such as a stratum total for a certain variable.

Consider a population total Y obtained by summing up individual values Y_{hij} for elements j over PSU i (giving the PSU total Y_{hi}), then over all PSUs in stratum h (giving the stratum total Y_h), and finally over all strata in the population:

$$Y = \sum_h Y_h = \sum_h \sum_i Y_{hi} = \sum_h \sum_i \sum_j Y_{hij}$$

The above is estimated by summing appropriately weighted values over the units in the sample. Firstly, the aggregate from a primary selection is estimated by a weighted sum of values for individual elements in the sample:

$$y_{hi} = \sum_j w_{hij} \cdot y_{hij} \tag{2.3}$$

These are then summed over primary selections and strata to obtain an estimate y for the population total Y:

$$y = \sum_h y_h = \sum_h \sum_i y_{hi} = \sum_h \sum_i \sum_j w_{hij} \cdot y_{hij} \tag{2.4}$$

The quantities w_{hij} are the weights associated with individual elements, determined in accordance with the estimation procedure. The basic or 'design' weights are taken as inversely proportional to the probabilities of selection of individual elements, but they may also involve other adjustments required in the process of estimation. The weights can be scaled such that y estimates Y, actually or within a constant scaling factor. For the present purpose, the scale is arbitrary and can be chosen as convenient. For instance in a self-weighting sample with a constant overall sampling rate $f = 1/F$, individual weights may be taken as unity (so that the total y estimates Y/F), or as F (so that y estimates Y), or as any other convenient constant. Similarly for a non-self-weighting sample, the individual weights may be scaled such that their average per sample element is 1.0, or equals a uniform inflation factor such as F.

In a similar manner, y_h estimates the stratum total Y_h . Each weighted PS value y_{hi} estimates the quantity (Y_h / a_h), as does their mean (y_h / a_h), where a_h is the number of PSUs selected at random in stratum h. The variance of individual PS estimates is estimated by the averaged squared quantities

$$var(y_{hi}) = \frac{1}{a_h - 1} \cdot \sum_{i=1}^{a_h} \left(y_{hi} - \frac{y_h}{a_h} \right)^2$$

and that of their total y_h estimated from a random sample of size a_h by:

$$var(y_h) = \frac{a_h}{a_h - 1} \cdot \sum_i \left(y_{hi} - \frac{y_h}{a_h} \right)^2 \tag{2.5}$$

Finally, with independent sampling across strata, we have

$$var(y) = var(\sum_h y_h) = \sum_h var(y_h) \tag{2.6}$$

The remarkable and convenient feature of the above expression is that it involves only the <u>appropriately weighted PS totals</u> y_{hi}, without explicit reference to the structure and manner of sampling within PSUs. This makes the variance estimation formula relatively simple, not requiring the computation of separate variance components in a multi-stage design. This also gives the method great flexibility in handling diverse sampling designs, and is indeed one of its major strengths and reasons for its widespread use in survey work.

<u>Paired Selections</u>

A particularly simple and useful special case of the above may be noted. With exactly two selections per stratum ($a_h = 2$), Equation (2.6) becomes

$$var(y) = 2 \cdot \sum_h \left[\left(y_{h1} - \frac{y_h}{2} \right)^2 + \left(y_{h2} - \frac{y_h}{2} \right)^2 \right] = \sum_h (y_{h1} - y_{h2})^2 = \sum_h \Delta y_h^2 \tag{2.7}$$

where Δy_h is the difference between the two PS values in stratum h.

In many surveys, stratification of the PSUs is carried out to a point where exactly two units are selected from each stratum, this being the minimum number required to estimate sampling errors. In practice the model is used even more widely for the following reasons. (i) Often it is considered desirable to stratify to the maximum extent possible, creating as many strata as the number of PSUs to be selected, or even more than that number using special techniques such as 'controlled selection'. (ii) Even more commonly, primary units are selected from ordered lists using systematic sampling, which can be regarded as implicit stratification with one unit selected per implicit stratum. In either of the above situations, variance computation requires redefinition or 'collapsing' of the strata to ensure that more than one sample PSUs is assigned to each stratum so redefined. The paired selection model provides the minimum collapsing necessary. (Section 4.4.)

The two-psus-per-stratum model is sometimes referred to as the Keyfitz method (Keyfitz, 1957), though 'paired selections' is the more commonly used term.

<u>The Finite Population Correction</u>

The above formulation is based on the assumption that sampling is with replacement at all stages. Usually sampling is done without replacement and the above expressions need to be modified to take that feature into account. This can be achieved in part using the concept of so-called 'ultimate clusters' (see Kalton, 1979, for a description). Suppose that the sample is selected with replacement at all stages except the last; at the last stage ultimate units are selected <u>without</u> replacement. We also assume that the overall sampling rate (f_h) is uniform within each stratum. Such a sampling scheme is equivalent to dividing the population exhaustively into what has been called 'ultimate clusters', till all elements in the population have been accounted for, and selecting a simple random sample (without replacement) of a_h ultimate clusters with rate f_h from each stratum h. The concept of ultimate clusters denotes - in the same way as primary selections - the aggregate of elements included in the sample from one selection of a PSU, and surmises the results of the series of operations involved in obtaining the final sample. With this model, the variance estimation formula becomes

$$var(y) = \sum_h \left[(1-f_h) . \frac{a_h}{a_h-1} . \sum_i \left(y_{hi} - \frac{y_h}{a_h} \right)^2 \right] \tag{2.8}$$

The above takes into account the effect of sampling without replacement at the <u>last</u> stage, on the assumption that sample elements have been selected with equal probability sampling of elements ('epsem'), throughout or within each stratum separately. Sampling without replacement at higher stages is not taken into account. Also, more complicated and approximate approaches may be required when element selection probabilities vary more generally. Kish (1965; p.432) proposes computing an effective sampling rate f as an appropriately weighted average of several f_g values which may exist in the sample for different components (g) of variance, the weights being proportional to the corresponding components of total variance. Usually approximate weights suffice for the purpose. In any case, in many national household surveys, neglecting the finite population correction is of no great practical consequence.

2.3 EXTENSION TO NON-LINEAR STATISTICS: VARIANCE OF RATIOS AND DIFFERENCES BETWEEN RATIOS

2.3.1 RATIOS

The combined ratio estimator of two aggregates y and x

$$r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j w_{hij} . y_{hij}}{\sum_h \sum_i \sum_j w_{hij} . x_{hij}} \tag{2.9}$$

is perhaps the most common statistic involved in survey analysis. Both the numerator (y) and denominator (x) may be substantive random variables - as for example in the estimation of income per capita from a household survey, where y is the total income and x the total number of persons estimated from the survey. Ordinary means, percentages and proportions are special cases of ratios, and therefore need not be discussed separately. In the case of a mean, the denominator is a count variable, ie x_{hij} is identically equal to 1 for all elements in the sample. This gives

$$\bar{y} = \sum_{hij} w_{hij} \cdot y_{hij} \;/\; \sum_{hij} w_{hij}.$$

For a proportion (percentage) the additional condition is that y_{hij} is a dichotomy equal to 1 (100) or 0 depending on whether or not unit j possesses the characteristic whose proportion (percentage) is being estimated.

Linearisation

Estimation of variance for a non-linear statistic with the primary selection comparison method requires linearisation of the estimator using Taylor approximation. For a ratio r = y/x this gives the well-known expression:

$$var(r) = \frac{1}{x^2}.[var(y) + r^2.var(x) - 2r.cov(x,y)] \tag{2.10}$$

where var(y) and var(x) are as defined above, and cov(x,y) is given by a similar expression:

$$cov(x,y) = \sum_h [(1-f_h).\frac{a_h}{a_h-1}.\sum_i (x_{hi}-\frac{x_h}{a_h}).(y_{hi}-\frac{y_h}{a_h})\;] \tag{2.11}$$

This extension of the method to a non-linear statistic requires some further assumptions in addition to [1]-[3] noted at the beginning of Section 2.2 above:

[4] The number of primary selections is large enough for valid use of the ratio estimator and the linearisation approximation involved in the standard expression for its variance.

[5] The quantities x_{hi} in the denominator (which often correspond to the sample sizes per PSU; henceforth referred to as the 'cluster sizes') are reasonably uniform in size within strata.

The last mentioned requirement is concerned with keeping the bias of the ratio estimator small. The relative bias in r is given approximately by the expression:

$$\frac{bias(r)}{r} = \left[\frac{var(x)}{x^2} - \frac{cov(x,y)}{xy}\right]$$

More important in practical terms is the observation that relative error of the denominator of the ratio, se(x)/x, provides an upper bound for the bias in r. According to Kish (1965), ideally this relative error should be below 0.1, and anyway should not exceed 0.2 when ratio estimation is used. Its value depends on the variability within strata of the PS sizes ('cluster sizes'), as well on the number of primary selections in the sample. It is an objective of well-designed samples to keep the variability in these sizes small. The problem can be more serious in estimates for subclasses of the population the selection of which cannot be fully controlled in the design of the sample.

Computational Simplification

A useful simplification is obtained by introducing the computational variable

$$z_{hi} = \frac{1}{x}.(y_{hi} - r.x_{hi}); \quad z_h = \sum_i z_{hi}; \quad z = \sum_h z_h = 0 \quad \textit{by definition.} \tag{2.12}$$

This reduces var(r) to the same form as var(y) of a simple total:

$$var(r) = \sum_h \left[(1-f_h).\frac{a_h}{a_h - 1}.\sum_i \left(z_{hi} - \frac{z_h}{a_h} \right)^2 \right] \tag{2.13}$$

For a detailed treatment of variances of ratios and their differences, see Kish and Hess (1959). A numerical example appears later in this section.

Computations for Subclasses

The above formulae can be used to compute variances of ratios estimated over subclasses of the sample: the procedure is simply to exclude from all summations any units not belonging to the subclass of interest. However, two types of complications can arise in moving from the total sample to subclasses.

[1] The appearance into the sample of many subclasses cannot be controlled. This can be the case in particular of subclasses defined in terms of rare or ill-distributed characteristics of individual elements, such as the level of education, occupation, or ethnic group of individual persons. Consequently, the denominators x (which generally refer to the weighted count of subclass cases selected in the PSUs) may become too variable, increasing the bias involved in the ratio estimation. In the extreme case (but by no means a rare one for very small or not so well-distributed subclasses), the subclass sample may be confined to a single PSU in some strata. Variance computation will then require (further) collapsing of the strata to ensure that at least 2 primary selections are available from each redefined stratum.

[2] Many subclasses of interest are confined to only a subset of PSUs in the sample. This applies to highly segregated classes, and especially to geographical domains (such as regions of a country), for each of which separate results may be required. With reduced number of primary selections available for computing sampling errors for such subclasses, the variability of the variance estimates is increased. (Section 4.4.5).

2.3.2 DIFFERENCES BETWEEN RATIOS

Comparisons between different sub-populations or between samples at different times is also a common objective of many surveys.

If the two ratios being compared

$$r = y/x, \quad and \quad r' = y'/x'$$

come from independent samples or strata, the variance of their difference is simply the sum of their individual variances.

However, in multi-stage designs the ratios being compared are usually estimated from sample elements coming from the same PSUs and their covariances must be taken into account. For the difference of two ratios

$$(r-r') = \frac{y}{x} - \frac{y'}{x'}$$

the standard expression for variance is

$$var(r-r') = var(r) + var(r') - 2.cov(r,r') \tag{2.14}$$

where var(r) and var(r') are as defined earlier, and

$$cov(r,r') = \frac{1}{xx'}.[cov(y,y')+r.r'.cov(x,x')-r.cov(y',x)-r'.cov(y,x')] \tag{2.15}$$

As before, all terms involved above are computed from sample values <u>appropriately weighted and aggregated to the PS level</u>.

Also by introducing

$$z_{hi} = \frac{1}{x}.(y_{hi}-rx_{hi}) - \frac{1}{x'}.(y'_{hi} - r'x'_{hi}) \tag{2.16}$$

the expression for var(r-r') can be greatly simplified and reduced to exactly the same form as that for var(y) of a simple aggregate:

$$var(r-r') = \sum_h [(1-f_h).\frac{a_h}{a_h-1}.\sum_i (z_{hi}-\frac{z_h}{a_h})^2]$$

ILLUSTRATION 2A NUMERICAL EXAMPLE OF THE COMPUTATIONAL PROCEDURE

The computational formulae above can be most clearly illustrated by considering a small sample with 2 PSUs per stratum (the paired selection model). The following example is based on Kish (1965, Sec. 6.5; also discussed in Kish, 1989, Chapter 13).

Quantities y_{h1} and y_{h2} are the weighted estimates from the two primary selections in stratum h, and $\Delta y_h = y_{h1} - y_{h2}$ their difference. Similar quantities are defined for variable x. The summations are taken over all strata h. Estimates of the two totals (y=149; x=255) are given in the totals row of Table 2A.(1), columns [1] and [4] respectively. With paired selection, their variances and covariances are given by the simple expressions (which also appears in the totals row of the table):

$$var(y) = \sum \left[y_{h1}-y_{h2}\right]^2 = \sum \Delta y_h^2 \quad (=217; col[3])$$

$$var(x) = \sum \left[x_{h1}-x_{h2}\right]^2 = \sum \Delta x_h^2 \quad (=475; col[6])$$

$$cov(x,y) = \sum \Delta x_h . \Delta y_h \quad (=293; col[7])$$

For the ratio $r = y/x = 149/255 = 0.58$, we define

$$z_{hi} = \frac{1}{x}.[y_{hi}-r.x_{hi}], \quad i=1,2$$

The result is shown in col[8]; note that by definition, the sum z = 0.

Next, var(r) can be computed from

$$var(r) = \frac{1}{x^2}.\left[\sum \Delta y_h^2 + r^2.\sum \Delta x_h^2 - 2r.\sum \Delta x_h.\Delta y_h \right]$$

$$= \frac{1}{255^2}\left[217 + 0.58^2 * 475 - 2 * 0.58 * 293\right] = 5.65 * 10^{-4} \quad \text{(i)}$$

or alternatively, making use of the computational simplification explained above, from

$$var(r) = \left[\sum \Delta z_h^2 \right] \quad \text{(ii)}$$

where the Δz_h values can be computed in either of the two forms:

$$\Delta z_h = z_{h1}-z_{h2} = \frac{1}{x}.\left[(y_{h1}-r.x_{h1}) - (y_{h2}-r.x_{h2})\right]$$

$$\Delta z_h = \frac{1}{x}.[\Delta y_h - r.\Delta x_h] = \frac{1}{x}.\left[(y_{h1}-y_{h2})-r.(x_{h1}-x_{h2})\right]$$

The two forms give identical results (col[9]). The totals row of col[10] gives var(r) computed according to eq.(ii).

Similarly, Table 2A.(2) gives the results for the second pair of variables (y' and x').

For the difference of two ratios

$$(r-r') = \frac{y}{x} - \frac{y'}{x'} \quad (=\frac{149}{255} - \frac{77}{156} = 0.09)$$

we have var(r-r') = var(r) + var(r') - 2.cov(r,r'). The two variance terms are computed as above, and cov as

$$cov(r,r') = \frac{1}{x.x'}.\left[\sum \Delta y_h.\Delta y'_h + r.r'.\sum \Delta x_h.\Delta x'_h - r.\sum \Delta x_h.\Delta y'_h - r'.\sum \Delta x'_h.\Delta y_h\right]$$

$$= \frac{1}{255*156}[\ 120 + 0.58*0.49*219 - 0.58*196 - 0.49*83\ ] = 6.96*10^{-4} \quad ...(iii)$$

This can be simplified for computations as

$$cov(r,r') = \sum [\Delta z_h.\Delta z'_h] \qquad \text{(iv)}$$

The results appear in the totals row of col[15]. Col[16] shows the quantities z_{hi}'' (adding to zero by definition). The quantities $\Delta z_h''$ can be computed from cols[9] of Tables 2A.(1) and (2), or from col[16] - ie, using either of the following two equivalent forms:

$$\Delta z''_h = \Delta z_h - \Delta z'_h = [z_{h1} - z_{h2}] - [z'_{h1} - z'_{h2}]$$

$$\Delta z''_h = z''_{h1} - z''_{h2} = [z_{h1} - z'_{h1}] - [z_{h2} - z'_{h2}]$$

Their identical result appears in col[16]. Written in terms of these quantities, the expression for var(r-r') is greatly simplified, though it does not show the contribution of its different components:

$$var(r-r') = \sum \Delta z''^2_h$$

The results are given in the totals row of col[18].

Note that for computational accuracy and convenience in Table 2A, quantities like z and their differences Δz are shown multiplied by the factor 10^2, this factor being the order of magnitude of the denominator total x; this makes the scale of quantities like z as used in the computation similar to that of the estimated ratio r = y/x. The square of these quantities, and hence var(r), are multiplied by 10^4. Note also that though the figures below have been printed to two decimal places, the actual computations were done to much higher accuracy.

TABLE 2A. ILLUSTRATION OF THE COMPUTATIONAL PROCEDURE.

(1) Variables x and y

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
11	11			19				-0.04		
2	9	2	4	16	3	9	6	-0.14	0.10	0.01
21	8			10				0.85		0.00
2	6	2	4	10	0	0	0	0.06	0.78	0.62
31	6			13				-0.63		0.00
2	15	-9	81	20	-7	49	63	1.30	-1.93	3.71
41	13			23				-0.17		0.00
2	5	8	64	8	15	225	120	0.13	-0.30	0.09
51	9			13				0.55		0.00
2	4	5	25	6	7	49	35	0.19	0.36	0.13
61	4			10				-0.72		0.00
2	7	-3	9	13	-3	9	9	-0.23	-0.49	0.24
71	5			7				0.36		0.00
2	7	-2	4	10	-3	9	6	0.45	-0.10	0.01
81	4			8				-0.26		0.00
2	5	-1	1	12	-4	16	4	-0.79	0.52	0.28
91	9			12				0.78		0.00
2	9	0	0	15	-3	9	0	0.09	0.69	0.47
01	9			20				-1.05		0.00
2	4	5	25	10	10	100	50	-0.72	-0.33	0.11
SUM=	149	7	217	255	15	475	293	0.00	-0.69	5.65
	y		var(y)	x		var(x)	cov(x,y)			var(r)

r = 0.58;
var(r) = 5.65*10^{-4} (Computed from equation (i)).

Column headings:

[1]= y_{hi}; [2]= Δy_h; [3]= Δy_h^2 in stratum h unit i; similarly [4]-[6] for x.

[7]= $\Delta x_h . \Delta y_h$; [8]= z_{hi}; [9]= Δz_h; [10]= Δz_h^2.

[11]= $\Delta x_h . \Delta y_h$; [12]= $\Delta x_h . \Delta x_h'$; [13]= $\Delta x_h . \Delta y_h'$; [14]= $\Delta x_h' . \Delta y_h$.

[15]= $\Delta z_h . \Delta z_h'$; [16]= z_{hi}''; [17]= $\Delta z_h''$; [18]= $(\Delta z_h'')^2$.

(Table continued)

Table 2A (cont.)

(2) Variables x' and y'

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
11	5			12				-0.59		
2	6	-1	1	9	3	9	-3	1.00	-1.59	2.53
21	1			1				0.32		
2	7	-6	36	13	-12	144	72	0.37	-0.05	0.00
31	2			10				-1.88		
2	9	-7	49	10	0	0	0	2.61	-4.49	20.13
41	7			12				0.69		
2	4	3	9	6	6	36	18	0.67	0.02	0.00
51	3			5				0.34		
2	1	2	4	6	-1	1	-2	-1.26	1.60	2.56
61	6			13				-0.27		
2	2	4	16	4	9	81	36	0.02	-0.28	0.08
71	3			6				0.02		
2	3	0	0	4	2	4	0	0.66	-0.63	0.40
81	0			1				-0.32		
2	4	-4	16	10	-9	81	36	-0.60	0.28	0.08
91	2			13				-2.83		
2	1	1	1	1	12	144	12	0.32	-3.16	9.96
01	10			18				0.71		
2	1	9	81	2	16	256	144	0.01	0.71	0.50
SUM=										
	77	1	213	156	26	756	313	0.00	-7.59	36.24
	y'		var(y')	x'		var(x')	cov(x'y')			var(r')

r' = 0.49
var(r') = $36.24 \cdot 10^{-4}$ (Computed from equation (i)).

(3) Computations for the difference of ratios

	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]
1 1					0.55			
2	-2	9	-3	6	-0.15	-1.14	1.69	2.85
2 1					0.52			
2	-12	0	0	-24	-0.04	-0.31	0.83	0.69
3 1					1.26			
2	63	0	49	0	8.64	-1.31	2.56	6.56
4 1					-0.86			
2	24	90	45	48	-0.01	-0.54	-0.32	0.11
5 1					0.21			
2	10	-7	14	-5	0.57	1.45	-1.24	1.54
6 1					-0.46			
2	-12	-27	-12	-27	0.14	-0.25	-0.21	0.04
7 1					0.33			
2	0	-6	0	-4	0.06	-0.20	0.54	0.29
8 1					0.05			
2	4	36	16	9	0.15	-0.19	0.24	0.06
9 1					3.61			
2	0	-3	-3	0	-2.17	-0.23	3.84	14.77
0 1					-1.77			
2	45	160	90	80	-0.23	-0.73	-1.04	1.08
SUM=								
	120	219	196	83	6.96	0.00	6.89	27.99
					cov(r,r')			var(r-r')

r-r' = 0.09
cov(r,r') = 6.96(Computed from equation (iii)).
var(r-r') = 27.99(Computed from the full expressions (i) and (iii)).

2.4 GENERALIZATION TO OTHER COMPLEX STATISTICS

The above are particular examples of the approach of linearisation of non-linear statistics using Taylor expansion.. In general terms, suppose the objective is to estimate the variance of a statistic which is a non-linear function of aggregates Y_1 to Y_s :

$$Z = f(Y_1, Y_2, \ldots, Y_s)$$

which is estimated from the sample by z defined in the same form

$$z = f(y_1, y_2, \ldots, y_s)$$

where the y's are sample aggregates estimating the corresponding population totals Y's. To terms of the first degree in (z-Z), the Taylor series approximation for z assumed close to (in the neighbourhood of) Z is

$$z = Z + \sum_k (y_k - Y_k).D_k; \quad D_k = \partial Z / \partial Y_k$$

where the partial derivatives D_k are evaluated at z=Z and taken as constants.

The above, with the added assumption that the unknown constants D_k can be replaced by their sample estimates d_k, gives

$$Var(z) = Var(\sum D_k.y_k) = Var(\sum d_k.y_k) = Var(z_L) \tag{2.17}$$

The above means that variance of non-linear z can be approximated by that of z_L, a <u>linear</u> function of the simple aggregates y_k,

$$z_L = \sum_k d_k.y_k \tag{2.18}$$

giving the general expression:

$$var(z) = var(\sum_k d_k.y_k) = \sum_k d_k^2.var(y_k) + \sum_{k \neq m} d_k.d_m.cov(y_k, y_m)$$

The above expression involves a <u>sXs</u> covariance matrix of simple aggregates y_k (k=1..s), with s variance terms and s(s-1)/2 pairs of identical covariance terms. These can be evaluated from expressions of the form given earlier for linear statistics. As an example, consider an ordinary ratio $z = y_1/y_2$, for which

$$d_1 = \partial z / \partial y_1 = 1/y_2;$$
$$d_2 = \partial z / \partial y_2 = -z/y_2^2;$$
$$var(z) = \frac{1}{y_2^2}.\left[var(y_1) + z^2.var(y_2) - 2z.cov(y_1, y_2)\right].$$

The expression above is exactly the same as the expression (2.10) given earlier for var(r), observing that here

$$y_1 = y, \quad y_2 = x, \quad and\ z = r$$

in terms of the notation used earlier.

Computational Simplification

The general expression for variance can be greatly simplified for computational purposes observing the following, based on Woodruff (1971), building on the work of Keyfitz (1957) and Kish (1968). In the linear statistic z_L, the aggregates y_k are by definition (appropriately weighted) sums of PS estimates

$$y_k = \sum_{hi} y_{k,hi}$$

which means that

$$z_L = \sum_k d_k \cdot y_k = \sum_k d_k \cdot (\sum_{hi} y_{k,hi})$$

Noting that d_k are constants not dependent on (h,i), the order of summation over k and (h,i) can be reversed giving

$$z_L = \sum_{hi} (\sum_k d_k \cdot y_{k,hi}) = \sum_{hi} z_{hi}, \ say$$

It follows that z_L can also be written as the simple aggregate of quantities computed at the PS level, namely of quantities

$$z_{hi} = \sum_k d_k \cdot y_{k,hi} \quad where \quad d_k = \partial Z / \partial Y_k \ at \ Z = z; \qquad z_h = \sum_i z_{hi} \tag{2.19}$$

resulting in the concise expression for var(z), without the need to work out the full covariance matrix:

$$var(z) = var(z_L) = \sum_h \left[(1-f_h) \cdot \frac{a_h}{a_h - 1} \cdot \sum_i (z_{hi} - \frac{z_h}{a_h})^2 \right]. \tag{2.20}$$

It only remains to develop the expressions for and numerically evaluate the partial derivatives of the estimate z from the sample data.

Particular Applications

Below are examples of particular applications of the above to various non-linear estimators (z) encountered in practical survey work. All that is necessary is to specify the quantities z_{hi} to be formed at the PS level in each case. To present the expression more concisely, the following abbreviated notation has been used:

$$t_{hi} = \frac{1}{x}.(y_{hi} - r.x_{hi}) = r.(\frac{y_{hi}}{y} - \frac{x_{hi}}{x})$$

In the notation, y, x, y', x' etc denote simple aggregates; y_{hi} etc the corresponding PS values, and z the statistic of interest such as a ratio $z = y/x$. Note that the PS values are appropriately weighted estimates; for example

$$y_{hi} = \sum_j w_{hij}.y_{hij},$$
$$giving \;\; y_h = \sum_i y_{hi}; \;\; y = \sum_h y_h. \tag{2.21}$$

The first two cases of the following have already been considered in more detail.

[1] Ordinary ratio

$$z = y/x; \quad z_{hi} = t_{hi}$$

[2] Difference of two ratios

$$z = r - r' = y/x - y'/x'; \quad z_{hi} = t_{hi} - t'_{hi}$$

[3] A weighted sum of ratios

$$z = \sum_k W_k.r_k = \sum_k W_k.(y_k/x_k) \; ; \quad z_{hi} = \sum_k W_k.t_{k,hi}$$

[4] Ratio of ratios ('double ratio')

$$z = r/r' = \frac{y}{x} / \frac{y'}{x'}; \quad z_{hi} = \frac{1}{r'}.[t_{hi} - t'_{hi}]$$

[5] Product of ratios

For the product of two ratios

$$z = r.r' = \frac{y}{x} . \frac{y'}{x'};$$
$$z_{hi} = r'.t_{hi} + r.t'_{hi} = z.\left(\frac{t_{hi}}{r} + \frac{t'_{hi}}{r'}\right)$$

More generally, for any product of ratios:

$$z = \prod_k r_k\,; \quad z_{hi} = z.\sum_k \left(\frac{t_{k,hi}}{r_k}\right).$$

[6] A weighted sum of double ratios

$$z = \sum_k W_k.\frac{r_k}{r'_k} = \sum_k W_k.z_k \ (say); \quad z_{hi} = \sum_k \frac{W_k}{r'_k}.[t_{k,hi} - z_k.t'_{k,hi}]$$

Double ratios appear for example when we consider ratio of indices (which are themselves ratios) for two periods - as in the relative of any current year which is the ratio of some index for that year to the index for some base year. Estimates of such relatives may be differenced, averaged or otherwise combined from different periods or samples, and so on. This gives rise to the general form [6]; forms [1] to [5] above are special cases of this. For a detailed treatment of standard errors of such indexes, see Kish (1968).

[7] A statistic such as the regression coefficient, expressed as

$$z = \sum_{hij} w_{hij}.y_{hij}.x_{hij} \,/\, \sum_{hij} w_{hij}.x_{hij}^2 .$$

This can be handled in a similar way by first defining at the level of individual elements (j) the following two variables

$$u_{hij} = y_{hij}.x_{hij}; \quad v_{hij} = x_{hij}^2$$

which can be aggregated over the PS's and strata in the usual way:

$$u = \sum_h u_h = \sum_{hi} u_{hi} = \sum_{hij} w_{hij}.u_{hij}$$

and a similar expression for v. This gives z in the form of the ratio u/v, so that

$$z_{hi} = \frac{1}{v}.(u_{hi} - z.v_{hi})$$

Tepping (1968) provides a general description of the approach and among other things, shows how variances of multiple regression coefficients may be estimated.

2.5 APPLICATION OF THE METHOD IN PRACTICE

Though the basic assumptions regarding the structure of the sample for application of the method are met reasonably well in many large-scale national household surveys, often the assumptions are not met exactly. Some of the more common approximations and how they may be dealt with in practice are noted below, with some examples. The issues will be discussed more fully in Chapter 4.

Systematic sampling.

Systematic sampling of primary units is a common and convenient procedure. Pairing of adjacent units to form strata to be used in the computations is the usual practice, resulting in the paired selection model noted earlier. Of course this results in some over-estimation of variance, to the extent the possible gain of ordering of the pair of units in each 'collapsed stratum' is disregarded. While this over-estimation is not avoided, variance of the variance estimator can be reduced (by a factor of 3/4) by using an alternative scheme of grouping of the units. This scheme is to utilize all (a-1) successive differences in the ordered list of a primary units, in place of only the a/2 comparisons among non-overlapping pairs. With this, the basic expression for variance becomes

$$var(y) = \frac{a/2}{a-1} \cdot \sum_{i=1}^{a-1} [y_i - y_{i+1}]^2. \qquad (2.22)$$

Deep stratification.

Often stratification is carried to a point where only one or even less than one primary unit is selected per stratum. This requires collapsing of similar strata to define new strata such that each contains at least two selections, which are then assumed to be independent. Such collapsing or grouping must be done on the basis of (similarity in) characteristics of the strata, and not of the particular units which happen to be selected. Otherwise the variance could be seriously underestimated.

Small and numerous primary selections.

Sometimes the primary units are too small, variable or otherwise inappropriate to be used directly in the variance estimation formulae. More suitable computational units may be defined by such techniques as random grouping of units within strata, and linking or combining of units across strata.

ILLUSTRATION 2B SOME EXAMPLES

Variance estimation from comparison among primary selections is perhaps the most widely used method in practical survey work. The following are just two examples which bring out some of the points noted in the previous section concerning the definition of the sample structure and computing units for application of the method. The following examples are based on two national surveys, in Sri Lanka and Malaysia, both conducted under the World Fertility Survey programme some years ago.

Example 1 (Sri Lanka).

The survey involved a two stage sample of households. At the first stage census enumeration areas (EAs) were selected systematically with probability proportional to population size from geographically ordered lists stratified by region, type of place and a hierarchy of administrative divisions within each region. From roughly 50 explicit strata, a sample of 700 PSUs was selected such that the resulting sample of households was self-weighting within each region. However, to compensate for non-response and other shortcomings in implementation, additional weights were introduced at the PSU level. The number of households selected per PSU was small, averaging around 10, but there was also a fair amount of variability around that average. By merging small PS's with others in the sample, around 400 somewhat larger and more uniform computing units were formed. Within each explicit stratum, adjacent newly defined units were paired to form around 200 'computing strata', and the two units within each such stratum were assumed to have been selected independently with replacement. (The overall sampling rate was small enough for the finite population correction to be negligible.) As a good practice, the survey data files contained codification of the strata and primary units as defined above for computational purposes, as well as the final estimation weights to be applied to individual elements in the sample. Most statistics of interest were in the form of combined ratios, for which the paired selection model could be applied over the 200 computing strata; an alternative would have been to use successive differences between adjacent units following their order of selection, separately within each of the 50 explicit strata.

Example 2 (Malaysia).

The survey population was divided into two sampling domains with very different designs. In the urban domain a single stage systematic sample of around 1500 dwellings was selected from geographically ordered lists. In rural areas a sample of around 4500 households was selected in two or three stages, starting with a sample of 100 localities at the first stage. For this purpose the rural sector was divided into 100 explicit strata, so as to select one locality (PSU) per stratum.

The construction of suitable computing strata and units (primary selections) required a number of steps. In the urban sector the single stage sample of dwellings was divided into 30 zones, each zone containing around 50 adjacent dwellings from the ordered list. These zones served as the computing strata. Within each zone, sample dwellings were allocated alternatively to form two computing units. Such a system is expected to reflect the actual systematic selection of dwellings.

In the rural domain, strata were examined on the basis of information available prior to sample selection, and paired on the basis of certain characteristics related to the subject matter of the survey. (Ideally, such pairing of strata should be done before sample selection to avoid subjective bias.) Each pair so defined constituted a computing stratum, and the sample of two localities within each gave the pair of primary selections for the purpose of variance estimation.

It may be mentioned that in practice, difficulties were encountered in computing sampling errors because information on characteristics of the original strata (not only the selected PSUs), required for their pairing, was not adequately documented. This brings out the importance of the often noted point that it is essential to preserve a full description of the sample structure, preferably as an integral part of the survey data files.

2.6 TECHNICAL NOTE ON THE BASIS OF THE METHOD

In a multistage design, each stage of sampling contributes to the overall variance of the survey estimate. Some survey practitioners may have the mistaken impression that computing variances simply from a comparison among estimates at the PSU level amounts to neglecting the contribution to variance of sampling at stages below the first stage of selection. While the derivation of the basic results of sampling theory is outside the scope of this Technical Study, the following discussion in the context of a simple two-stage design is presented to clarify the basis of the method described in this chapter. Of course, the results are available in more detail in many good text books on sampling. (See for example, Kish 1965, Section 5.6.)

To understand the basics, we consider a population of clusters of equal size (B), with a two stage design consisting of a random selection of (a out of A) clusters, followed by the selection of a constant number (b out of B) of elements within each selected cluster. The overall sampling rate f = ab/AB = n/N is constant for all elements in the population.

Firstly, it can be seen that for a population divided into clusters, the variability between elements in the population

$$\sigma^2 = \sum_{ij} (Y_{ij} - \bar{Y})^2 / AB$$

can be decomposed into two components:

[1] the between-cluster component

$$\sigma_a^2 = \sum_i (\bar{Y}_i - \bar{Y})^2 / A$$

[2] and the within-cluster component

$$\sigma_b^2 = \sum_{ij} (Y_{ij} - \bar{Y}_i)^2 / AB$$

giving

$$\sigma^2 = \sigma_a^2 + \sigma_b^2$$

In the context of survey sampling, we use slightly different quantities

$$S^2 = \frac{N}{N-1}.\sigma^2; \quad S_a^2 = \frac{A}{A-1}.\sigma_a^2; \quad S_b^2 = \frac{B}{B-1}.\sigma_b^2$$

in terms of which the above decomposition can be approximated as:

$$S^2 \approx S_a^2 + \frac{B-1}{B}.S_b^2.$$

on the reasonable assumption that N and A (respectively the number of elements and PSUs in the population) are large.

With the two stage design imposed on this population, the two components of variance of an estimate such as the sample mean are:

[1] the between-cluster component

$$Var(\bar{y})_a = (1-\frac{a}{A}).\frac{S_a^2}{a} \tag{2.23}$$

[2] and the within-cluster component, which for a sample of b elements from a single cluster would be

$$Var(\bar{y})_{b,1} = (1-\frac{b}{B}).\frac{S_b^2}{b}$$

so that for the actual sample of a clusters it becomes

$$Var(\bar{y})_b = (1-\frac{b}{B}).\frac{S_b^2}{a.b}. \tag{2.24}$$

Total variance is the sum of the two components:

$$Var(\bar{y}) = (1-\frac{a}{A}).\frac{S_a^2}{a} + (1-\frac{b}{B}).\frac{S_b^2}{a.b} \tag{2.25}$$

To estimate these components from the sample, we define similar quantities (with the summations over the sample)

$$s_a^2 = \sum_i (\bar{y}_i-\bar{y})^2/(a-1) \tag{2.26}$$

and

$$s_b^2 = \sum_{ij} (y_{ij}-\bar{y}_i)^2/[a(b-1)]. \tag{2.27}$$

The second quantity is based on a simple random sample within each cluster, and hence provides an unbiased estimate of the corresponding population value for the within-cluster component:

$$E(s_b^2) = S_b^2 \tag{2.28}$$

However, the first quantity is not simply an estimate of the between-cluster component: it reflects that variability as well as the additional variability resulting from the fact that it is based only on a subsample of elements within each cluster. It can be shown that its expected value is given by the following.

$$E(s_a^2) = S_a^2 + (1-\frac{b}{B}).\frac{S_b^2}{b}. \tag{2.29}$$

Substituting into Equation (2.25), it is seen that an approximately unbiased estimate of variance is provided by

$$var(\bar{y}) \approx (1-\frac{ab}{AB})\frac{s_a^2}{a} = (1-f).\frac{s_a^2}{a}, \tag{2.30}$$

the approximation involved in the above expression being :

$$\frac{a}{A}.(1-\frac{b}{B}).\frac{s_a^2-s_b^2/b}{a}$$

which is small, especially in national samples where the first stage sampling rate (a/A) is small.

The implication of the above result can be generalized to many practical multi-stage designs involving complex subsampling within PSUs of variable sizes. Therefore, it is a result of great practical utility; it indicates that good estimates of the total variance can be provided simply by certain sample quantities aggregated to the level of PSUs, <u>without explicitly involving any reference to the complexity of subsampling within the primary units</u>. Of course the subsampling does influence efficiency of the resulting sample, but its effect is largely incorporated in the s_a^2 term computed from the sample values. With variable cluster sizes and sample takes, generally sample aggregates (rather than cluster means per element) are estimated first, and then proportions or means etc are estimated as ratios of the relevant aggregates. For these, variances are given by expression like (2.8) and (2.13).

Some discussion on variance components in the context of a general, more complex sampling design is given in Section 5.6 below.

3

COMPUTING SAMPLING ERRORS: COMPARISON AMONG SAMPLE REPLICATIONS

3.1 INTRODUCTION

The idea of replicated variance estimation was introduced in Section 2.1. In the linearisation method described in the preceding chapter, the primary selections are taken as the replicates and their sampling variability computed within each stratum separately, and then aggregated across the strata. The alternative approach is to consider replications of the full sample, each of which is of the same design and reflects full complexity of the sample, including its stratification, and provides a valid estimate of the statistic of interest. A replication differs from the full sample only in sample size. However, to be useful in the present context, the size of each replication needs to be large enough:

[1] for it to reflect the structure of the full sample, and

[2] for any estimate based on a single replication to be close to the corresponding estimate based on the full sample.

The approach involves division of the full sample, by design or subsequent to selection in some manner, into a set of replications each of the same design. Using the same estimation procedure, estimates y_j for each replication, their average $\bar{y}$(Eq. 2.1), and the corresponding estimate from the full sample are obtained. The form of the estimators can be of any complexity: it may for example be a population total, a ratio, a combination or function of ratios, regression or correlation coefficients etc; it may relate to the total population, or to any subclass of the population distributed across the replications. We will denote the full sample estimate as y for the linear case, and as $\tilde{y}$ for the non-linear case.

The replications are normally constructed such that among them they cover all the units in the sample, each unit once or the same constant number of times. This means that for a linear statistic - ie an estimate which is a linear function of the sample values - the replicated average $\bar{y}$ is the same as the full-sample estimate y. However, for non-linear estimators (which are the main reason for using the method), the two are not necessarily the same. This is illustrated by the following simple example. For the ratio y of two sample aggregates x_1 and x_2, we have the estimate from replication j

$$y_j = \frac{x_{1j}}{x_{2j}}$$

giving the average of replicated estimates as

$$\bar{y} = \frac{1}{n}.\sum y_j = \frac{1}{n}.\sum \frac{x_{1j}}{x_{2j}}$$

while the full-sample estimate is

$$\tilde{y} = \frac{x_1}{x_2} = \frac{\sum x_{1j}}{\sum x_{2j}} .$$

The two are not the same, since the former is the average of separate ratios while the latter is a combined ratio.

The main attraction of the replication comparison method to variance estimation arises from the fact that, with sufficiently large and complex replications each reflecting the full sample, the estimates y_j from individual replications, and even more so their average $\bar{y}$, are expected to be close to the same estimate based on the full sample. This is the case exactly in relation to a linear statistic such as a sample total (y), but approximately also the case for a non-linear estimate ($\tilde{y}$) of any complexity based on the full sample. This also applies to the closeness between the variance of the simple average of replicated values, $var(\bar{y})$, and $var(\tilde{y})$ of the full-sample estimate of interest. This means that if the former can be obtained, it can be used to approximate the latter, irrespective of the complexity of the estimate $\tilde{y}$. Practical application of this very convenient approach has two requirements, in addition to [1] and [2] noted above regarding the size and structure of each replication to reflect the full sample, namely:

[3] that a procedure is established to estimate $var(\bar{y})$, which is then used to approximate $var(\tilde{y})$;

[4] and that the number of replications used is sufficiently large to yield a variance estimate with adequate precision.

Within the group of methods based on comparison among replications, two quite distinct approaches may be noted:

A. a comparison among estimates based on independent replications which together comprise the full sample; and

B. a comparison among estimates based on overlapping 'pseudo' or repeated replications, constructed by repeated resampling of the same parent (full) sample.

Method [A] is based on the assumption that the parent sample can be regarded as consisting of a number of independent replications or subsamples, each reflecting the full complexity of the parent sample, differing from it only in size. With these assumptions the replicated estimates can be regarded as independent and identically distributed (IDD) random variables, so that the variability among them gives in a very simple form, a measure of variance of the overall sample estimator. The limitation of the method is that in many situations the total sample cannot be divided into a sufficient number of independent replications of adequate size for the method to be applicable; its strength is its simplicity when applicable. The method is discussed with illustrations in Section 3.2.

Method [B] refers to the family of resampling methods for computing sampling errors for complex designs and statistics in which the replications to be compared are generated through repeated resampling of the same parent sample. Each replication is designed to reflect the full complexity of the parent sample. In contrast to the independent replication method whose limitations the resampling methods are designed to overcome, the resampling methods are based on overlapping replications which reuse the sample selections in several (many) computing units. With repeated resampling the variance estimates are made more stable through averaging over many subsamples. However, since the replications generated are not independent, special procedures are required to control the bias in the variance estimates provided by their comparison.

Various resampling procedures have been developed which differ in the manner in which replications are generated from the parent sample and the corresponding variance estimation formulae evoked. Three general procedures known as the balanced repeated replication (BRR), jackknife repeated replication (JRR), and the bootstrap are available, though the last is not yet established for general use in the presence of complex selection methods. Generally, the resampling methods can more easily deal with complex statistics and estimation procedures in comparison with other methods. However, they tend to be technically and computationally more complex; they are also somewhat more restrictive in the sample designs handled in comparison with the linearisation approach of Chapter 2. In relation to statistical properties of the variance estimates produced, the three methods - linearisation, BRR and JRR - have been found to yield comparable and generally satisfactory results in complex situations (Section 4.3). The basic difference between JRR and BRR may be stated as follows. With JRR, a replication is formed by dropping a small part of the total sample, such as a single PSU in one stratum; consequently each replication measures the contribution of a small part such as a single stratum. In BRR, a replication is formed by dropping a part (such as one half) of every stratum and it measures the variance of the entire sample. Comparing BRR and JRR, the former generally requires less computational effort, but can have two major disadvantages: it is technically more complex; and more importantly, it tends to be more restrictive in the type of sample designs handled. For these reasons, JRR is the preferred approach. However, a certain lack of readily available general purpose computer software is a disadvantage common to the resampling methods at present.

3.2 VARIANCE FROM INDEPENDENT REPLICATIONS

3.2.1 THE PROCEDURE

The basic requirement of the method is that, by design or subsequent to sample selection, it should be possible to divide the parent sample into more or less independent replications, each with essentially the same design as the parent sample. In a multistage design, for example, the parent sample has to be divided at the level of the PSUs, ie divided exhaustively into a number of non-overlapping replications each consisting of a separate independently selected (and ideally also independently enumerated) set of PSUs. For a combined estimation over a number of strata, each replication must itself be a stratified sample covering all the strata.

With independent replications, each providing a valid estimate of the same population parameter of interest, the results of the theory of 'independent replicated variance estimator' noted in Section 2.1 can be directly applied. The variance of the simple average of n replicated estimates y_j

$$\bar{y} = \frac{\sum_j y_j}{n}; \quad var(\bar{y}) = \frac{1}{n}.\left[\frac{\sum_j (y_j-\bar{y})^2}{n-1}\right] \tag{3.1}$$

provides an estimate of the variance of the same estimate from the full sample, exactly for a linear estimate y, or approximately for a non-linear estimate $\tilde{y}$. A somewhat conservative estimator (giving higher value) is obtained by replacing $\bar{y}$ in the summation by $\tilde{y}$; that is by writing:

$$var(\tilde{y}) = \frac{1}{n}.\left[\frac{\sum_j (y_j-\tilde{y})^2}{n-1}\right] \tag{3.2}$$

The above may be modified to incorporate the finite population correction if that is important. With a uniform overall sampling rate f for the full sample, this amounts to inserting the factor (1-f) on the right hand side. With variable sampling rates, a simple approach would be to take "(1-f)" above as an appropriately averaged value (see Section 2.2).

3.2.2 CONSTRUCTING INDEPENDENT REPLICATIONS

It is useful to begin by stating the requirements which should be met ideally in application of the procedure. The basic requirement is that the parent (full) sample is composed of a number of independent subsamples or replications, each with the same design and procedures, but selected and implemented independently. The requirement of common and independent procedures applies to sample selection as well as to data collection and estimation.

1. Sample Selection. The replications should be designed according to the same sample design, on the basis of the same frame and type of units, system of stratification, sampling stages and selection methods, etc. as the parent sample. In drawing several replications from the same population, independence requires that each replication is replaced into the frame before the next is drawn and the randomised selection procedure is applied separately for each selection.

2. Data Collection. Following sample selection, the procedure for data collection should be the same and applied independently for each replication. Data collection refers to various steps in the whole measurement process, including questionnaire design, staff recruitment and training, mode and procedures for data-collection, fieldwork organisation, supervision and control, recording and coding of responses, data entry, and so on. Independent application of common data collection procedures requires, for example, that independent sets

of field staff (supervisors, interviewers, coders, etc) drawn in principle from a common pool, are used for different replications.

3. Estimation. A common estimation procedure refers not only to the mathematical form of the particular estimator used, but also to all the other steps involved in computing the final estimates from the survey data - steps such as data editing, imputation, treatment of outliers, weighting and other adjustments. Independent application means that all steps in the estimation procedure are applied separately to each replication. For example, if the sample data are weighted to agree with certain population control totals, it is implied that the relevant weights are determined independently for each replicated subsample, as distinct from using a common set of weights determined on the basis of the full data set.

3.2.3 APPLICATION IN PRACTICE

Approximation to Independent Replications

In practice the above requirements are rarely met exactly. For instance, if the estimation procedure is complex, repeating all its steps for each replication can be too expensive and time consuming. Hence while separate estimates are produced for each replication (as must be done for the variance estimation procedure to be applied), some steps in the procedure - such as imputation, weighting, and adjustment of the results against external control totals - are applied only once to the sample as a whole. The results can be different if these steps were applied to the sample results from each replication separately and independently. Strict independence of data collection procedures is even more difficult to implement. That would require organisation and implementation of numerous steps in the measurement process independently for each replication, possibly involving a great increase in cost and inconvenience. (Some such separation in an appropriate form can of course be useful in the assessment of non-sampling errors.)

Perhaps the most critical requirement is that of independent selection of replications following the same design. Ideally, the full sample may be formed by combining independent subsamples. Usually, however, it is a matter of partitioning an existing sample into more or less independent subsamples. It is important to note that in a multistage design, the partitioning of the sample should be done at the level of primary selection, ie all sample elements within a PS should be assigned to the same replication. To estimate the total variance across strata in a stratified sample, each replication must itself be a stratified sample parallelling the parent sample.

The sample may be divided into replications at the time of selection or subsequently, after selection. Consider for instance a systematic sample in which primary units are to be selected with interval I; the sample may be selected in the form of n replications, each selected systematically with a distinct random start and selection interval = (n.I). A more convenient and common alternative is to select the full sample in one operation, but in such a way that it can be subsequently divided into subsamples which are by and large independent and reflect the design of the full sample. As an example, consider a systematic sample of 500 PSUs to be divided into 20 replications each of size 500/20 = 25 PSUs. One may imagine an ordered list of the 500 sample units as divided into 25 'zones', each comprised of 20 adjacent units. A replication would consist of one unit taken from each zone: for instance the first unit from each zone forming the first replication, the second unit from each zone forming the second replication, and so on. In fact, such a simple scheme can be applied with great flexibility and permits many straightforward variations. The units in the full sample may have been selected with uniform or varying probabilities: the above subsampling scheme retains the original relative probabilities of selection. If the original sample is stratified, one may order the selected units stratum after stratum and divide the entire list into equal zones for the application of the above procedure. The effect of original stratification will be reflected in the replications if the number of units to be selected is large enough for all or most strata to be represented in each replication. Alternatively, units may be cross-classified by zone and stratum, ie each stratum divided into a number of zones and each zone linking units (sample PSUs) across a number of strata. Deming (1960) provides many examples and extensions of such procedures.

Choice of the Number of Replications

In most multistage designs the number of primary selections involved is limited, which constrains the number of replications into which the sample may be divided. There are of course samples in which the total number of PSs available is so inadequate that the number of replications and the number of units per replication both have to be rather small. In that situation the method of independent replication is inappropriate for variance estimation. However, when the sample design permits, choice still has to be made between the extremes of having many small replications, or having only a few but large replications. If many replications are created, the number of PSs per replication may become too small to reflect the structure of the full sample. This will tend to bias the variance estimation. On the other hand variances estimated from only a small number of replications tend to be unstable, ie themselves subject to large variance. There is no agreement as to the most appropriate choice in general terms. Kish (1965, Section 4.4), for example, summarises the situation as follows: "Mahalanobis (1946) and Lahiri (1958) have frequently employed 4 replicates... Tukey and Deming (1960) have often used 10 replicates... Jones (1956) presents reasons and rules for using 25 to 50 replicates. Generally I too favour a large number perhaps between 20 and 100."

The primary argument in favour of having many replications (each necessarily comprised of a correspondingly small number of units) is that the variance estimator (equation 3.1) is more precise and the statistic $\bar{y}$ (average of the replicated estimates) is more nearly normally distributed. The precision of the variance estimator decreases as the number of replications is reduced. Furthermore, for a given value of variance or standard error, the interval associated with any given level of confidence becomes wider. This is because with n replications, estimate $var(\bar{y})$ is based on (n-1) "degrees of freedom", and in constructing the confidence interval for the population parameter in the form

$$\bar{y} \pm t.[var(\bar{y})]^{1/2}$$

the value of t (and hence the width of the above interval) corresponding to any <u>given</u> level of confidence increases with decreasing n. This can be seen from the standard Student-t distribution. Another consideration is that with a small number of replications, it is necessary to assume that the individual replicated estimates y_j are normally distributed, though mild departures from normality are generally not important; fortunately the assumption of normality of y_j is improved as the number of primary selections per replication is increased. In any case, when the number of replications is large, it is necessary to assume only that the <u>mean</u> $\bar{y}$ is normally distributed.

On the other hand, having fewer and larger replications also has some statistical advantages. (1) With large sample size per replication, the individual replicated estimates y_j are more stable and more nearly normally distributed. This helps in inference. (2) The replicated estimates y_j, and even more so their average $\bar{y}$ is closer to the estimate $\hat{y}$ based on the full sample for non-linear statistics as well. This facilitates extension of the method of variance estimation to non-linear statistics, which is the main justification for its use. (3) Most importantly, increasing the number of primary units per replication makes it easier to reflect the structure of the full sample in each replication, which reduces the bias in the variance estimator.

Also should be noted some practical considerations and wider objectives for opting for a smaller number of replications, each consequently larger and potentially more complex in design: (4) With fewer replications, there is less disturbance of the overall design as a result of the need to select the sample in the form of independent replications. (5) The additional cost and difficulty involved in separate measurement and estimation is smaller. (6) With a smaller number of replications, it is more feasible to appropriately randomise the work allocation of interviewers, coders, etc to measure non-sampling components of variance. (7) Replicated or 'interpenetrating' designs can be useful for more general checking of survey procedures and results. These objectives are better served when the number of replications to be dealt with is small. Lahiri (1957) for instance provides a number of illustrations from the Indian National Sample Survey. (8) The same is true of displaying the survey results separately by replication to convey to the user a vivid impression of the variability in sample survey results (Illustration 7D). It is of course also possible to have a larger number of replications for more stable sampling error estimation, and collapse them to a smaller number for objectives (6), (7) and (8).

In view of the above conflicting considerations and opinions, it is not possible to make specific recommendations on the appropriate choice of the number and size of replications. With say 100-1000 PSUs in the sample, a simple rule which has been found rather reasonable is to begin by making both the number of replications and the number of sample PSUs per replication equal to the square-root of the given total number of PSUs in the sample. For example, with somewhat over 200 sample PSUs, it would result in around 15 replications each with 15 PSUs. Similarly, with 600 or so PSUs, one would begin by considering around 25 replications, each with 25 PSUs (Illustration 3A).

ILLUSTRATION 3A SOME EXAMPLES OF REPLICATED SAMPLING

The following provide a number of examples of the actual or potential use of replicated sampling in practical survey work, especially in developing countries.

<u>Example 1. Samples with Several Hundred PSUs Selected Systematically</u>

One of the basic factors favouring more wide-spread use of replicated sampling in developing countries is that, in many situations, various practical considerations favour the use of sample designs with relatively small and compact but numerous PSUs. The stratification is often largely obtained through systematic selection of PSUs from geographically ordered lists, rather than through an elaborate system with many explicit strata. These features of the design facilitate the division of the sample into a number of independent replications, each with a reasonably large number of sample PSUs to reflect the structure of the full sample. For instance many surveys are based on samples of census enumeration areas (EAs). Typically these are quite compact areas with say 50-300 households on the average, from which say 10-50 households may be selected into the sample per EA. Hence national surveys typically based on samples of several thousand households may include several hundred EAs as primary units. An example is provided by the national fertility surveys conducted under the World Fertility Survey during 1972-84: the following were among the surveys based on 300 or more PSUs each, even though the sample sizes were modest (mostly in the range 4000-7000 households per survey; see Verma 1980; Scott and Harpham, 1987):

Country	No.of PSUs in sample	Country	No.of PSUs in sample
Ghana	300	R.of Korea	390
Portugal	300	Peru	410
Pakistan	326	Jamaica	428
Panama	354	Venezuela	480
Senegal	358	Trinidad	648
Indonesia	366	Philippines	742
Colombia	370	Sri Lanka	750

In most cases the sample areas were selected systematically after stratification and ordering of the lists by type of place, administrative division and geographical location. Following the same ordering, the full sample can be easily divided into replications systematically, each replication essentially retaining the original stratification. For instance the 480 PSUs in Venezuela may be systematically partitioned into say 12 replications of 40 areas each, taking unit numbers 1, 13, 25, etc, into the first replication, unit numbers 2, 14, 26, etc, into the second, and so on; or one may construct 24 replications each with 20 areas; or some other combination of the number and size of replications. If areas in the original sample were selected with PPS (probability proportional to size), the PPS character of the selection will be retained in the replications as well. Estimates for any statistic, however complex, may be produced for each replication separately, and variance of the estimator for the full sample computed simply by an expression of the form (3.1)

Example 2. A Replicated Master Sample

An example covering many household surveys on diverse topics from one country is provided by the sampling scheme developed in relation to the sampling frame created from the 1990 Population Census of Indonesia. (The following description is based on unpublished documentation at the Indonesia Central Bureau of Statistics.) From the Census frame of areas, a large master sample of around 4000 EAs (PSUs) is selected with PPS after urban-rural, administrative and geographical stratification. Retaining the original ordering of selection, the combined master sample list (formed by placing one stratum after another) is systematically divided into around 50 replications, each with 80 or so EAs. Subsampling from the master sample for a survey on any particular topic would generally involve selecting simply a subset of the replications at random, followed by listing and sampling of households within the selected areas to yield a sample with the required number of areas and households. Depending on the topic, a national sample may contain say 400-2000 EAs, ie 5-25 replications. In practice the system is more flexible than may appear because systematic division of the ordered master sample lists can be easily repeated with any subsampling interval. For instance by doubling the interval, twice as many replications can be created each with half as many EAs; the same result may be obtained by systematically selecting one in two areas from each original replication. Such modifications can be convenient for smaller surveys. In a similar fashion, fewer and larger replications may be constructed for bigger surveys. The attractiveness of the scheme for the routine production of approximate estimates of sampling errors from household surveys covering diverse topics is obvious. There is scope for improvement in the precision of variance estimates through averaging over related surveys. The main requirement in estimating variances would be to tabulate the survey results separately by replication.

Example 3. Independent Enumeration of Subsamples

In the Indian National Sample Survey (NSS), the division of the total sample for any annual round into two or more (most often two) subsamples within each explicit stratum has been a permanent feature of the design. Presentation of the final estimates by subsample along with the full-sample estimates help in conveying to the user a rough but vivid idea of the degree of uncertainty involved in the survey results. Often, unexpectedly large divergences between the subsample estimates can help to locate exceptional field or data processing problems. (This indeed was a major objective of the approach.) Furthermore, estimates of standard errors can be computed at stratum as well as the total level. On certain assumptions about distribution of the subsample estimates, non-parametric confidence intervals can be constructed for the population parameters. In some earlier rounds of the NSS, different field staff surveyed different subsamples, and tabulation of the subsamples was done at different processing centres. In this manner, comparison of subsample results provided an indication not only of the sampling error, but also of some components of variance arising from other, nonsampling, sources (India, 1990). Some of the results, showing the range of estimates from different replications, have been reproduced in Illustration 7D.

Example 4. Replications in a Single Stage Stratified Sample

The use of replicated sampling can of course be most convenient and appropriate in single stage designs with a large number of 'primary' units. Such is the case with many surveys of establishments where individual establishments form the units of selection and analysis. Many surveys of households in developed countries involve direct sampling of households from lists in a single stage. In developing countries as well, there are examples of household surveys using single stage designs at least in selected domains such as major urban centres. Though the following example is based on surveys of economic establishments rather than of households, it provides a good illustration of the use of replicated sampling.

The system of economic surveys in Cyprus consists of a set of annual surveys each covering a major sector of the economy (manufacturing, trade, transport, services, etc). These are supplemented by a monthly survey of employment covering all the sectors. The surveys are based on a common design. The population of establishments is divided into economic sectors, sectors into subsectors with varying degrees of detail, and within each subsector a single stage sample of establishments is selected with PPS (probability proportional to the size of employment) systematically from establishments ordered by size of employment. A wide range of estimates is required from the survey separately for a number of domains (economic sectors or subsectors), covering many variables such as annual sales, gross output, direct costs, value added, investment and employment - both as aggregates and as averages per establishment and per worker. In addition, monthly levels and trends in employment by sector are produced using a complex composite estimation procedure applied to the employment survey (Cyprus, 1990).

To compute variances, the sample of establishments within each domain was divided systematically into a number of replications. (A domain usually referred to a subsector at the 2 digit level of ISCO.) Replications could be formed simply and flexibly because of the systematic nature of the sample within each domain. The main issue requiring consideration was the choice of the appropriate number of replications in each domain. An adequate member of replications per domain and an adequate number of units per domain are both important requirements. The domains varied considerably in sample size (mostly in the range of 50-400 establishments), and the following simple rule was used to determine the number of replications in a uniform way: the number of replications and the number of units per replication both varied in proportion to the square-root of sample size in the domain. Thus in a domain with 100 sample establishments, 10 replications each with 10 establishments were created; in a domain with 300 sample establishments, around 17 replications each with 17 or so establishments were created. (Some results on coefficients of variation computed by using the independent replication approach are shown in Illustration 6.E.(2); these were obtained from past annual surveys for the purpose of sample redesign for future surveys.)

3.3 JACKKNIFE REPEATED REPLICATION (JRR)

3.3.1 INTRODUCTION

The Jackknife Repeated Replication (JRR) method is one of the 'resampling methods' for computing sampling errors for complex designs and statistics in which the replications to be compared are generated through repeated resampling of the same parent sample. Each replication is designed to reflect the full complexity of the parent sample. However the replications in themselves are not independent but overlap, as their construction involves repeated resampling from the same parent sample in a specified manner. This general approach is useful and necessary when the independent replications are not available, or if their number is too small to yield useful estimates of sampling error using the simple replicated approach of Section 3.2. With repeated resampling the variance estimates are made more stable through averaging over many subsamples. However, since the replications generated are not independent, the simple expression (3.1) cannot be used to estimate the variance of a statistic based on the full sample, irrespective of whether the statistic of interest is linear or more complex. Also, special procedures are required in constructing the replications so as to control the bias in the variance estimates resulting from the lack of independence of the replications. Various resampling procedures are possible depending upon the manner in which the replications are generated from the parent sample and the corresponding variance estimation formulae evoked.

The JRR method involves the following basic steps:

[1] the selection of a number of overlapping subsamples from the parent sample;

[2] derivation of the needed estimates of the population from the subsamples;

[3] and an estimation of the variance of the parent sample estimator from the variability among the subsample estimates.

In the <u>basic</u> model of the JRR method, replications are generally formed by randomly eliminating one sample PSU from a particular stratum at a time, and duplicating or reweighting the retained PSUs in the stratum concerned to appropriately compensate for the eliminated unit. Hence with $\underline{a}$ primary selections in the full sample, the same number of unique replications are defined, each corresponding to a particular unit $\underline{i}$ in stratum $\underline{h}$ having been eliminated, and the other units in the stratum given appropriately increased weights so that the estimate $y_{(hi)}$ from the replication has the same expected value as the estimate y from the full sample. (However the two generally do not have identical values in any particular sample.) This is because each unit is eliminated or retained (with appropriately increased weight) in the construction of the replications exactly the same number of times. For the same reason, the average of the replicated estimates involving eliminations from the same stratum

$$y_{(h)} = \sum_i y_{(hi)}/a_h$$

and the average of all a replications

$$\bar{y} = \sum_{hi} y_{(hi)}/a; \quad where \quad a = \sum_h a_h$$

both actually (not merely in expectation) equal the total sample estimate in the linear case, and approximately so in the non-linear case. (In the above a_h is the number of sample PSUs in stratum h, and also the number of corresponding replications generated.)

With units eliminated from one stratum at a time in the construction of the replications, each replication provides a measure of only the variance contributed by the particular unit and stratum involved. These estimates are then aggregated over the replications to obtain the total variance.

The basic model can be generalised in various ways. For example, replications may be constructed by eliminating several PSUs at a time from a particular stratum. It is also possible to think of suitable jackknife procedures which leave out parts of more than one stratum at a time. The BRR approach discussed in Section 3.4, in which replications are formed by eliminating one half of every stratum in the sample at a time thus becomes a limiting case of the generalised JRR approach. Tukey (1968) for example sees possible advantages in the intermediate approach in which JRR replications are formed by eliminating anything less than half the sample at a time (Section 3.3.4).

The following description, however, is largely in terms of the usual JRR application of dealing with one unit in one stratum at a time.

3.3.2 DESCRIPTION OF THE PROCEDURE

The Method in the Linear Case

Consider a replication formed by dropping a particular PSU i in stratum h and appropriately increasing the weight of the remaining (a_h-1) PSUs in that stratum to compensate for the missing PSU. The estimate for a simple aggregate (total) for this replication is

$$\begin{aligned} y_{(hi)} &= (y - y_h) + \frac{a_h}{a_h-1}.(y_h - y_{hi}) \\ &= y - \frac{a_h}{a_h-1}.(y_{hi} - \frac{y_h}{a_h}) \end{aligned} \qquad (3.3)$$

The last term in (3.3) vanishes by definition if an average is taken over all i in the stratum. This means that in the linear case, and with each PSU dropped only once, the average of estimates $y_{(hi)}$ over the stratum

$$y_{(h)} = \sum_i y_{(hi)}/a_h \tag{3.4}$$

and the average over all $\underline{a} = \Sigma a_h$ replications

$$\bar{y} = \frac{\sum_h \sum_i y_{(hi)}}{\sum_h a_h} \tag{3.5}$$

both equal to total y estimated for the full sample. It also follows from the above definitions that the standard expression for variance of the total y estimated from the full sample

$$var(y) = \sum_h [(1-f_h).\frac{a_h}{a_h-1}.\sum_i (y_{hi}-\frac{y_h}{a_h})^2] \tag{3.6}$$

can be written in any of the following three forms (which are equivalent in view of the identity of (3.4), (3.5), and the full sample estimate y in the linear case):

$$var_1(y) = \sum_h [(1-f_h).\frac{a_h-1}{a_h}.\sum_i (y_{(hi)}-y_{(h)})^2] \tag{3.7}$$

$$var_2(y) = \sum_h [(1-f_h).\frac{a_h-1}{a_h}.\sum_i (y_{(hi)}-\bar{y})^2] \tag{3.8}$$

$$var_3(y) = \sum_h [(1-f_h).\frac{a_h-1}{a_h}.\sum_i (y_{(hi)}-y)^2] \tag{3.9}$$

<u>Extension to the Non-linear Case</u>

In the JRR method the standard variance form (3.6) is replaced by one of the three expressions (3.7)-(3.9); usually the last of the three is used, as it is more conservative. In the linear case this replacement makes no difference; but then there is no point in introducing the latter more complicated expressions. The point of introducing them is that they provide good approximations for the variance of more complex, non-linear statistics as well. This is because, being based on nearly the full sample, estimates like $y_{(hi)}$, $y_{(h)}$, and even more so their overall average $\bar{y}$ are expected to be close to the full-sample estimate $\tilde{y}$ for a complex statistic. Hence their variance, expressed by any of the last three forms, provides a measure of variance of $\tilde{y}$ as well. This is not true of the standard simpler form (3.6). In the non-linear case, expressions (3.7)-(3.9) are rewritten by replacing y (which we use to denote the full-sample estimate in the linear case) everywhere by $\tilde{y}$ (which denotes the same in the non-linear case).

Paired Selections

The case of exactly two sample PSUs per stratum ($a_h=2$) is a common and convenient one. In application, this amounts to eliminating one PSU and duplicating the other in a stratum at a time. To write the formulae for this special case simply, a slightly different notation is more convenient: $y'_{(h)}$ is used to denote the estimate formed by dropping a particular PSU from stratum h, and $y''_{(h)}$ to denote its complement formed by dropping the other PSU in the stratum. With this notation we can write the above as

$$var_1(\tilde{y}) = \frac{1}{4}\sum_h [(1-f_h).(y'_{(h)}-y''_{(h)})^2] \qquad (3.10)$$

or, by replacing the average of the two replications by their near-equivalent total sample estimate $\tilde{y}$, as

$$var'_1(\tilde{y}) = \sum_h [(1-f_h).(y'_{(h)}-\tilde{y})^2] \qquad (3.11)$$

or its complement

$$var''_1(\tilde{y}) = \sum_h [(1-f_h).(y''_{(h)}-\tilde{y})^2] \qquad (3.12)$$

With paired selection and using either of the above equivalent forms, it is possible to reduce (to half) the number of replications needed by only considering one primary selection at random from each stratum and disregarding its complement. This can result in considerable saving in computational work, and may not result in much loss in precision in large samples with many strata.

Other forms are also possible, such as from (3.9):

$$var_3(\tilde{y}) = \frac{1}{2}.\sum_h (1-f_h).\left[(y'_{(h)}-\tilde{y})^2 + (y''_{(h)}-\tilde{y})^2 \right] \qquad (3.13)$$

which turns out to be simply the average of (3.11) and (3.12) and hence requires twice as many replicated estimates as either of the two.

3.3.3 DEFINING THE SAMPLE STRUCTURE

The basic requirement for the JRR method is that the (full) sample be selected by dividing the survey population into a number of strata, from each of which two or more primary selections are obtained independently at random. In practical application of the method, various steps are often required to redefine or simplify the given sample structure to conform with the required model, and also for computational convenience and efficiency. This is the case especially when the number of primary units and strata involved is large, and the PS's tend to be small or variable in size. The steps, described more fully in Section 4.4, include: (i) random grouping of PS's within strata to form more suitable computing units; (ii) combining of units across strata; (iii) collapsing (disregarding) some stratification to ensure that at least two effective primary units are available from each computing stratum; (iv) treating adjacent units in a systematic sample as independent selections within the strata so defined; and (v) assuming that the primary selections within each stratum are independent.

As mentioned above, the basic idea of the JRR is to drop a random set of PS's from a stratum at a time. This means that in the computational formulae described, it is such 'drop-out groups' rather than individual PS's which are relevant. Of course, often such 'groups' consist of single PS's; nevertheless the conceptual distinction remains, and it is more convenient to think of the former as the effective computing units. We can therefore add to the above list of sample redefinition the following: (vi) random grouping of units further if necessary, to define 'drop out' groups which serve as the effective computing units, so that a replication is formed by dropping one such unit from a particular stratum; and (vii) other possible variations in the method to reduce the number of replications involved. These include for instance: considering only a subset of replicates for dropping, always retaining the others; dropping replicates from several strata at the same time; or permitting some units to be dropped out in more than one replication. However the most common (and arguably the most efficient) scheme is when sample PSUs are dropped one at a time and each unit is dropped exactly once.

3.3.4 WIDER USES OF THE JACKKNIFE APPROACH

In conclusion it is worth quoting the following remarks by Tukey (1968):

"One important point about the use of the jackknife - in which, rather than leaving out half of the available data, one leaves out smaller pieces in turn until all has been left out once - is its ability to be used at two or more levels. If one had used the jackknife method rather than the half-sample method [the reference is to the BRR method discussed in the next section] to obtain the DEFF or DEFT values [design effect; see Chapter 5], as in Kish and Frankel situation [see references at the end of this document], one could go ahead to estimate the stability of these results, or of their differences, or their ratios. By doing this we would have a better understanding of what these results, as well as many others, really mean.

"The technique is simple in principle, but often not easily grasped without detailed exposition. The basic idea in dealing with a DEFT, for example, would be to lay aside one piece of the data and then calculate the DEFT by jackknifing the remainder. This jackknifing would involve leaving out additional pieces of the data, one at a time, and in turn. Once this has been done for one first-stage piece, we proceed to do all this over again and again, laying aside each piece of the data at the first stage, we are ready to jackknife the DEFTs thus obtained and thus estimate their variability...It seems to me that there will, in the two-psu-per-stratum situation faced by Kish and Frankel, prove to be real advantages to a suitable jackknife procedure - one that leaves out more than one PSU, but less than half of all PSUs. If we have five strata, each with two PSUs, the half-sample method requires leaving out one PSU in each stratum, which can be done in 32 ways. A probably sensible jackknife approach would involve leaving out one PSU in each of, say, two of the five strata. There are 40 possible ways to do this. The gain will come from leaving out enough, but noticeably less than half of the data...

"A simple example on which to compare "jackknifing" and "halving" is the problem of data gathered in several blocks with three values, equally spaced in time, obtained in each block. This sort of data arises naturally in many agricultural problems (including time of planting and time of harvesting). Yates (private communication) suggested that, where the number of blocks was a power of two, we treat such situations by halving the data and comparing the halves, repeating this in an interesting and ingenious way according to a fractional factorial pattern, thus obtaining the full number of degrees of freedom for the variability estimate.

"Analysis of this problem shows that the bias due to halving - both in the location of the optimum data and in the estimate of the variance of this optimum data - is noticeably larger for halving than for 'leaving out one' jackknifing, which also provides the full number of degrees of freedom for a variance estimate. I believe we can expect to find this phenomenon rather general. Accordingly, I believe that 'leave out a few' techniques will do even better than halving in the two-psu-per-stratum situation."

3.4 BALANCED REPEATED REPLICATION (BRR)

In the Balanced Repeated Replication method, a replication is formed from the full sample by randomly selecting some and dropping the remaining units from every stratum. (Typically, a replication is composed of a random half of every stratum.) Consequently any replication when compared with the full sample (or with the average of all replications considered), provides a measure of the variance of the entire sample. These measures are then averaged over the whole set of replications to obtain more stable estimates.

The BRR method is technically more complex than the JRR; consequently, the discussion in this section needs to be more elaborate. Technical complexity can in fact be a serious drawback of the method when compared with JRR, though the two methods have been found to perform equally well in dealing with complex statistics under complex designs.

3.4.1 BASIC APPROACH OF THE BRR METHOD

The Linear Case

To illustrate the basic approach we begin by assuming the following model:

[1] The population is divided into a number of strata ($h = 1$ to H) and from each stratum exactly two independent primary selections are obtained.

[2] The objective is to estimate variance of a linear statistic such as a population total.

Assumption [1] refers to the structure of the total sample, which is then divided into a number of overlapping replications as described below. This assumption is not an unrealistic one for the BRR method, because the procedure is most readily applied to designs with 2 PSUs per stratum. Though it can be extended to 3 or more (but a constant number of) primary selections per stratum, it remains a fact that the method is not so flexible in dealing with diverse designs. However this limitation is not as restrictive as it may appear. (i) Firstly, many surveys do use 2 PSUs per stratum designs; or more commonly, such a design is approximated by applying the collapsed stratum technique to systematic samples or to designs with fewer than 2 PSUs per stratum. (ii) Secondly, through random grouping of units or combining across strata, the actual design can be redefined to fit the 2 PSUs per stratum design with good approximation (see Section 4.4 for further discussion of these issues). (iii) Thirdly, of course there is no restriction on the manner in which the sample is selected within primary units.

Assumption [2] is not realistic because the point of using a method like the BRR is to deal with complex, non-linear statistics. However, as in the case of the JRR method, the linear case provides a starting point for description of the method. Moreover, approximations for the non-linear case are obtained by analogies to the results for the linear case.

Suppose that a sample with two primary selections from each of H strata is divided into two parts as follows. One of the 2 PS's from each stratum is assigned at random to "Subsample 1", and the other PS to its complement "Subsample 2". Let y_h' be the appropriately weighted estimate of the stratum total from the unit from stratum h in Subsample 1, and y_h'' be the estimate from the unit in Subsample 2. Their average $y_h = (y_h' + y_h'')/2$ also estimates the stratum total. On the lines of (2.8), the ordinary estimator of variance of the full sample estimate

$$y = \sum_h y_h = \frac{1}{2} \sum_h (y_h' + y_h'') \tag{3.14}$$

is given by

$$var(y) = \sum_h (y_h' - y_h)^2 = \sum_h (y_h'' - y_h)^2 = \frac{1}{4} \sum_h (y_h' - y_h'')^2 \tag{3.15}$$

(Note that in the present notation, a quantity like y_h' is scaled to be twice the size of a 'half-sample estimate' like y_{hi} used in Illustration 2A.)

Alternatively we can consider the simple replicated estimator of variance following (2.2). The two samples form independent replications, their respective estimates of the population total being

$$y' = \sum_h y_h' \; ; \quad y'' = \sum_h y_h''$$

and their average

$$\bar{y} = \frac{1}{2}(y' + y'') = \frac{1}{2}\cdot\sum_h (y_h' + y_h'')$$

is identical to y in (3.14) for the linear case being considered. Following (2.2), the simple replicated variance of y is

$$var(y)_1 = (y'-y)^2 = (y''-y)^2 = \frac{1}{4}\cdot(y'-y'')^2 \tag{3.16}$$

Estimator (3.15) is based on H degrees of freedom, but (3.16) only on one degree of freedom (hence the subscript '1'). The latter therefore is much less precise than the former; however it is no more biased as can be seen by rewriting (3.16) as follows:

$$var(y)_1 = \frac{1}{4}\left[\sum_h (y_h'-y_h'')\right]^2 = \frac{1}{4}\cdot\sum_h (y_h'-y_h'')^2 + \frac{1}{4}\cdot\sum_{h\neq k}\left[(y_h'-y_h'')(y_k'-y_k'')\right]$$

The second term on the right vanishes in expectation, because the units are allocated to one or the other subsample independently across strata. This leaves the first term which is identical to (3.15).

The reason for considering a replicated estimator of the form of (3.16) is that, unlike the ordinary estimator (3.15), it can be readily extended to non-linear statistics $\bar{y}$ as will be explained below. Its major drawback is the lack of precision. To improve stability, the operation of creating 'half-sample replications' can be repeated many (T) times and average taken, giving the averaged estimator

$$y_T = \frac{1}{T}\cdot\sum_t y_t = \frac{1}{2T}\cdot\sum_t (y_t' + y_t'')$$

and its variance

$$var(y)_T = \frac{1}{4T}\cdot\sum_t (y_t'-y_t'')^2 \tag{3.17}$$

where subscript t refers to a particular half sample replication and its complement. It can be shown that if the average is taken over all possible half samples, (3.17) is as precise as (3.15). However, the number of possible half samples is too large (= 2^{H-1} distinct half samples and as many complements) for the above to be useful in practice. Instead, a much smaller "balanced" set of replications is sought which can achieve the precision of (3.15). To explain the idea of balancing (McCarthy, 1966), it is illuminating to express the totals (y_t', y_t'') estimated from a particular replication (and its complement) in terms of the quantities (y_h', y_h'') corresponding to arbitrarily defined but fixed Subsamples 1 and 2 referred to earlier. It can be easily verified that the relationship is

$$(y_t'-y_t'') = \sum_h d_{t,h}\cdot(y_h'-y_h'')$$

where $d_{t,h}$ is an index defined such that:

$d_{t,h}$ = +1 if the unit appearing in replication t in stratum h is from Subsample 1, and
$d_{t,h}$ = -1 if the unit is from Subsample 2.

The above gives (noting that $d_{t,h}^2 = 1$ in all cases)

$$(y'_t - y''_t)^2 = \sum_h (y'_h - y''_h)^2 + \sum_{h \neq k} \left[(y'_h - y''_h)(y'_k - y''_k)(d_{t,h} \cdot d_{t,k}) \right]$$

Substituting into (3.17) and reversing the order of summations over t and h gives

$$var(y)_T = \frac{1}{4} \cdot \sum_h (y'_h - y''_h)^2 + \frac{1}{4T} \cdot \sum_{h \neq k} \left[(y'_h - y''_h)(y'_k - y''_k) \sum_t (d_{t,h} \cdot d_{t,k}) \right]$$

The second term on the right represents additional variability in comparison with the ordinary estimator (3.15). For all possible replications, this term vanishes for linear estimates (though only approximately for non-linear estimates, for which the method is needed and used). The idea of balancing is to choose a much smaller set which has the property that for every fixed pair of strata (h,k) the quantity

$$\sum_t d_{t,h} \cdot d_{t,k} = 0 \tag{3.18}$$

this making the additional variability disappear.

Another desirable property of the set is that the average y_T (eq. 3.17) is the same as the total y estimated from the full sample. This is ensured if in each stratum, the two units appear in the same number of replications, which requires that for every stratum h :

$$\sum_t d_{t,h} = 0 \tag{3.19}$$

<u>Extension to Non-linear Statistics</u>

On the assumption that the distribution of the average y_T of the replications is close to the distribution of the non-linear estimator $\bar{y}$ based on the full sample, equation (3.17) provides a good estimate of $var(\bar{y})$ as well. This applies irrespective of the complexity of the estimator. Hence the method is directly extended from simple linear statistics to non-linear statistics of any complexity, provided that the assumptions noted above remain valid. Several empirical investigations confirm this validity. It is also assumed that the method of constructing balanced replications (as described in Illustration 3B below) can be carried over from the linear case, for which it is established, to the non-linear case. Since in the non-linear case the replicated estimates y_t or their average y_T are not identical to $\bar{y}$ based on the full sample, the different forms shown in (3.16) are also not identical to each other. This gives the following four alternative estimators for the variance of $\bar{y}$:

$$v_1' = \frac{1}{T} \sum_t (y_t' - \bar{y})^2$$

$$v_1'' = \frac{1}{T} \sum_t (y_t'' - \bar{y})^2$$

$$v_2 = \frac{1}{2T} \sum_t \left[(y_t' - \bar{y})^2 + (y_t'' - \bar{y})^2 \right] = \frac{1}{2}(v_1' + v_1'')$$

$$v_3 = \frac{1}{4T} \sum_t (y_t' - y_t'')^2$$

Of the above four forms,the last one (v_3) is the same as (3.17). Note that while v_1' and v_1'' are each based on T replications, v_2 and v_3 involve T complementing pairs (ie 2T replications), and hence considerably more computational work. The forms v_1' and v_1'' are simply complements of each other, and the choice among them is arbitrary. Form v_2, being an average of the two, can be somewhat more precise but involves twice as many replicated estimates; it may be preferred if the computational work is no problem. Being a more conservative estimator, v_2 may also be preferred over v_3. In situations where the BRR method is appropriate, the difference between the last two should be small in any case.

3.4.2 APPLICATION IN PRACTICE

The BRR method is most conveniently applied to designs with two primary selections per stratum; extension to more than 2 (but a constant number) of PS's per stratum is possible though at the cost of further increase in complexity of the procedure for constructing the 'balanced set' as described in Illustration 3B below. This makes the method somewhat restrictive. In any event, appropriate redefinition of the sample structure may be necessary following the various approaches described in Chapter 4.

As explained in Illustration 3B, the number of replications required is between (H+1) and (H+4), where H is the number of strata, corresponding to $\underline{a}$ = 2H primary selections with the paired selection model. Consequently, the number of replications required is roughly half that required in the basic model for the JRR method (which requires as many replications as the number of primary selections). Also, the size of each replication in BRR is roughly half as large as that in JRR. For both these reasons, the computational work involved in the former is generally less than the latter, though in either of the two cases it usually exceeds that involved in the linearisation method.

In any case, in samples with many PSUs and strata, the computational costs of the BRR may become excessive. One way to reduce the cost is to appropriately combine units and strata to obtain fewer computing units and strata as discussed in Chapter 4. Another procedure is to reduce the number of replications required for a given number of strata by seeking only a partially balanced design. McCarthy (1966) describes a method of partial balancing in which the full set of strata is divided into a number of equal groups. A balanced set of part-replications is created for only one group, and each replication is completed for the whole sample by simply repeating the pattern of that group in all the other groups of strata. McCarthy also notes that some other authors found the same efficiency by selecting a random set from the full set of orthogonally balanced replications.

TABLE 3B.(1). Balanced set of 8 replications.
(Source: Kish and Frankel, 1970)

ORTHOGONAL BALANCE OF 8 REPETITIONS FOR 7 STRATA[a]

k	h: 1	2	3	4	5	6	7
1	+	+	+	−	+	−	−
2	−	+	+	+	−	+	−
3	−	−	+	+	+	−	+
4	+	−	−	+	+	+	−
5	−	+	−	−	+	+	+
6	+	−	+	−	−	+	+
7	+	+	−	+	−	−	+
8	−	−	−	−	−	−	−

[a] The first repetition is marked $+ + + - + - -$ in accord with the scheme in [28, p. 323] for creating orthogonal repetitions. It is an arbitrary representation of a random first choice from the pair of replicates in each stratum; $- - - + - + +$ would represent the associated complement replication. The repetitions from 2 to $(k-1)$ are designated from the first row by moving to the right one place circularly in Columns 1 to $(k-1)$. The kth repetition is all $-$. Note that the number of $+$ and $-$ replicates used are 4 for each stratum; also that the number of changes is 4 from any repetition to any other.

The situation is similar when k is any integral multiple of 4 and the number of strata is $H = k-1$. If $H = k-2$ or $H = k-3$, orthogonal balance may be obtained by omitting any 1 or 2 columns. If $H = k$, orthogonal balance may be obtained by writing a whole column of $-$ for the last stratum, using the same replicate from it for every repetition, but this sacrifices the symmetrical use of all replicates.

ILLUSTRATION 3B CONSTRUCTING A BALANCED SET OF REPLICATIONS

Plackett and Burman (1946) provide a method of constructing 'orthogonal' [T x T] matrices with entries +1 or -1 and T any multiple of 4, which satisfy equations (3.18) and (3.19) when summed over rows t. Such a matrix can be used to define a balanced set of half-samples as follows. To start with, in each stratum one unit is assigned at random to Subsample 1, and the other unit to Subsample 2. Let us identify each unit in the first set with a '+', and in the second set with a '-'. The orthogonal matrix consists of a '+' or '-' in each cell according to a certain pattern. The T rows of the matrix represent the set of T replications (half-samples), and any subset of H (out of T) columns represents the strata. A row defines the composition of a replication; ie it specifies the particular unit (a '+' or a '-') taken from each stratum to form the replication. The matrix has the property that equation (3.18) can be satisfied if $H \leq T$, but (3.19) only if $H<T$. Hence T can be taken as the next multiple of 4 after H; that is, the number of replications required for a sample with H strata is in the range T = (H+1) to (H+4).

<u>Examples</u>

Tables 3B.(1) to (3) provide examples of 'balanced' sets for T = 8, 16 and 24 replications respectively. The first example is taken from Kish and Frankel (1970). The rows define 8 half-sample replications for a sample with 7 (or fewer) strata; strata are identified by columns. It is assumed that each stratum contains 2 primary selections, one assigned a '+' and the other a '-' at random. Replication No. 1 for example is formed according to the first row, ie by taking units designated as '+' in strata 1, 2, and 3 and 5; and taking units designated as '-' in the other strata. Similarly Replication No. 8 takes the units designated as '-' in each stratum. It can be seen that in any column (stratum) the number of + and - signs is equal, meaning that each of the 2 units appears in the same number of replications. This satisfies (3.19). Also if any two columns are taken, the number of rows in which they have the same sign (giving $d_{t,h} . d_{t,k} = 1$) is equal to the number of rows in which the two columns have different signs (giving $d_{t,h} . d_{t,k} = -1$). This satisfies (3.18). Note that if there are fewer than 7 strata, the number of columns not required can be dropped arbitrarily. Similar remarks apply to the other two tables, which deal with somewhat larger samples. The footnote in Table 3B.(1) describes the symmetry of the pattern.

TABLE 3B.(2). Balanced set of 16 replications.

(Source: Plackett and Burman, 1946)

+ + + + − + − + + − − + − − −

The complete design is generated by taking this as the first column (or row), shifting it cyclically one place fourteen times and adding a final row of minus signs, thus:

+	−	−	−	+	−	−	+	+	−	+	−	+	+	+
+	+	−	−	−	+	−	−	+	+	−	+	−	+	+
+	+	+	−	−	−	+	−	−	+	+	−	+	−	+
+	+	+	+	−	−	−	+	−	−	+	+	−	+	−
−	+	+	+	+	−	−	−	+	−	−	+	+	−	+
+	−	+	+	+	+	−	−	−	+	−	−	+	+	−
−	+	−	+	+	+	+	−	−	−	+	−	−	+	+
+	−	+	−	+	+	+	+	−	−	−	+	−	−	+
+	+	−	+	−	+	+	+	+	−	−	−	+	−	−
−	+	+	−	+	−	+	+	+	+	−	−	−	+	−
−	−	+	+	−	+	−	+	+	+	+	−	−	−	+
+	−	−	+	+	−	+	−	+	+	+	+	−	−	-
−	+	−	−	+	+	−	+	−	+	+	+	+	−	−
−	−	+	−	−	+	+	−	+	−	+	+	+	+	−
−	−	−	+	−	−	+	+	−	+	−	+	+	+	+
−	−	−	−	−	−	−	−	−	−	−	−	−	−	−

TABLE 3B.(3). Balanced set of 24 replications.
(Source: McCarthy, 1966)

	Stratum																				
Half-Sample	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
1	+	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+	-	+	-	+	+
2	+	+	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+	-	+	-	+
3	+	+	+	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+	-	+	-
4	+	+	+	+	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+	-	+
5	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+	-
6	-	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+
7	+	-	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+	+	-	-	+
8	-	+	-	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+	+	-	-
9	+	-	+	-	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+	+	-
10	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+	+
11	-	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+	-	+	-	-	+
12	-	-	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+	-	+	-	-
13	+	-	-	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+	-	+	-
14	+	+	-	-	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+	-	+
15	-	+	+	-	-	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+	-
16	-	-	+	+	-	-	+	+	-	+	-	+	+	+	+	+	-	-	-	-	+
17	+	-	-	+	+	-	-	+	+	-	+	-	+	+	+	+	+	-	-	-	-
18	-	+	-	-	+	+	-	-	+	+	-	+	-	+	+	+	+	+	-	-	-
19	+	-	+	-	-	+	+	-	-	+	+	-	+	-	+	+	+	+	+	-	-
20	-	+	-	+	-	-	+	+	-	-	+	+	-	+	-	+	+	+	+	+	-
21	-	-	+	-	+	-	-	+	+	-	-	+	+	-	+	-	+	+	+	+	+
22	-	-	-	+	-	+	-	-	+	+	-	-	+	+	-	+	-	+	+	+	+
23	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+	-	+	-	+	+	+
24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Specification of Balanced Sets

Wolter (1985) provides a complete list of T X T orthogonal matrices for T as a multiple of 4 up to 100. The reader may consult that reference, if available. However to make this Technical Study as self-contained as possible, below we specify the set (except for T = 92 and 100 for brevity) in an alternative and much more concise form drawn from the original paper of Plackett and Burman (1946). It may be pointed out that matrices meeting the requirements (3.18) and (3.19) are not unique. For example the '+' and '-' signs can be interchanged; or rows and columns can be rearranged in any arbitrary way; or for H<T, any arbitrary subset of columns not required can be dropped. In Table 3B.(4), the matrices are specified in several forms:

[1] The most common presentation gives a single set of (T-1) values. (In the table, N has been used by the original authors in place of T here.) The full matrix is constructed from the given pattern as follows. Assume that the given pattern forms the top (T-1) entries of the first column of the matrix. Each next column up to (T-1) is formed by shifting each entry in the preceding column one row down, in a circular fashion, the (T-1)th row entry moving to the top. The final row and final column of minus signs are added.

[2] Some matrices are formed by doubling others. This follows from the property that if [T] is a matrix of the required type, then that is also the case for the following matrix of order 2T where [-T] means a matrix formed from [T] with all signs reversed.

$$\begin{bmatrix} T & T \\ T & -T \end{bmatrix}$$

[3] In a few cases (T = 28, 52 and 76; also 100 not shown here) the cyclic permutation is applied to blocks of cells rather than to individual cells as in the case of [1].

[4] Regarding the two cases not shown (T = 92 and 100) the pattern for the next higher T value can be used. One may for instance use T = 96 in a case actually requiring only T = 92; and for 100 use 104 obtained by doubling the pattern for 52 according to [2].

TABLE 3B.(4). Constructing balanced sets.

$N = 8$. + + + − + − −

$N = 12$. + + − + + + − − − + −

$N = 16$. + + + + − − − + + − − + − − −

$N = 20$. + + − − + + + + − + − + − − − − + + −

$N = 24$. + + + + + − + − + + − − + + − − − + − + − − − − −

$N = 28$. First nine rows

+ − + + + + − − −	− + − − − + − − +	+ + − + − + + − +
+ + − + + + − − −	− − + + − − + − −	− − + + + − + + −
− + + + + + − − −	+ − − − + − − + −	+ − + − + + − + −
− − − + − + + + +	− − + − + − − − +	+ − + + + − + − −
− − − + + − + + +	+ − − − − + + − −	+ + − − + + + + −
− − − − + + + + +	− + − + − − − + −	− + + + − + − + +
+ − + − − − + − +	− − + − − + − + −	+ − + + − + + + −
+ + + − − − + + −	+ − − + − − − − +	+ + − + + − − + +
+ + + − − − − + +	− + − − − − + − −	− − + − + + + − +

$N = 32$. − − − − + − + − + + + − + + − − − + + + + + − − − + − + − − +

$N = 36$. (Obtained by trial) − + − + + + − − − + + + + + − + + + − − + − − − − + − + − + + − − − + −

$N = 40$. Double design for $N = 20$.

$N = 44$. + + − − − + − + − − + + + − + + + + + − − − + − + + + − − − − − + − − − + + − + − + + −

$N = 48$. + + + + + − + + + + − − + − + − + + + − − + − − + + − + + − − − + − + − + + − − − − + − − − −

$N = 52$.

+ + −

First eleven rows					
+	− + − − − − − − − −	+ + − − − − + + + +	+ − + + − − − + − −	+ + − − + + − − + +	+ + + + + + − − − −
−	+ + − + − + − + − +	+ − − + − + + − + −	+ − + − − + + − − +	+ − − + + − − + + −	+ − + − + − − + − +
+	− − − + − − − − − −	+ + + + − − − − + +	− − + + + + − − + +	+ − − + − − + + − −	− − + + + + + + − −
−	− + + + − − − + − +	+ − + − − + − − + −	− + + − + − − + + −	+ − + − − + + − − +	− + + − + − + − − +
+	− − − − − + − − − −	+ + + + + + − − − −	+ + − − + + + + − −	− − + + + + − − + +	− − − − + + + + + +
−	− + − + + + − + − +	+ − + − + − − + − +	+ − − + + − + − − +	− + + − + − − + + −	− + − + + − + − + −
+	− − − − − − − + − −	− − + + + + + + − −	− − + + − − − + + +	+ + − − + + + + − −	+ + − − − − + + + +
−	− + − + − + + + − +	− + + − + − + − − +	− + + − − + + − + −	+ − − + + − + − − +	+ − − + − + + − + −
+	− − − − − − − − − +	− − − − + + + + + +	+ + − − + + − − + +	− − + + − − + + + +	+ + + + − − − − + +
−	− + − + − + − + − +	− + − + + − + − + −	+ − − + + − − + + −	− + + − − − + − + −	+ − + − − + − + + −

$N = 56$. Double design for $N = 28$.

$N = 60$. + + − + + + − + − + − − + − − − + + − + + + + − − + + + + + − − − − − + + − − − − + − − −
+ + − + + − + − + − − − + −

$N = 64$. Double design for $N = 32$.

$N = 68$. + + − − + − + − − + + − − − + + + + − + − + + − + + + − − + − − − + − + + + − + + − − − −
− − + − + − − − − + + + − − + + − + − + + −

$N = 72$. + + + + + + + − + + + − + − − + + − + + + − − − + + − + − + + − + − − − + + + − + − − + −
+ − − + + + − − − + − − + + − + − − − + − − − − − −

$N = 76$.

+	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −
+	− +	− −	+ +	− −	− −	+ +	+ +	− −	+ −	− −	− −	− −	− −	+ +	+ +	+ +	− −	− +	+ +	+ +	+ +	− −
−	+ +	− +	+ −	− +	− +	+ −	+ −	− +	+ −	− +	− +	− +	− +	− +	+ −	+ −	+ −	− +	+ −	+ −	+ −	− +

+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −	+ −
+ +	+ +	+ +	− −	− −	− −	− −	+ +	− −	+ +	+ +	− −	− −	+ −	− −
+ −	+ −	+ −	− +	− +	− +	− +	+ −	− +	+ −	+ −	− +	− +	+ −	− +

The first three rows are given; to obtain the complete design the square blocks are permuted cyclically. The first column, apart from the corner element, has alternate signs.

$N = 80$. + + + − + + − − + + + + − + − − + − + + + + + + − + + − − − − + + − − − + − + − + − + − +
+ + − − + + + + − − + − − − − − − + − + + − + − − − − + + − − + − −

$N = 84$. + + − + + − − + − + + + + − − − + + − − − + − + − + + + + + + + − + − − + + + − + + − − +
+ − − + + − + − − − − − − − + − + − + + + − − + + + − − − − + − + + − − + −

$N = 88$. Double design for $N = 44$.

$N = 92$. This design has not yet been obtained.

$N = 96$. Double design for $N = 48$.

4

COMPUTING SAMPLING ERRORS IN PRACTICE

4.1 PREREQUISITE: MEASURABILITY

The most basic requirement for generating information on sampling errors is that the survey design and procedures adopted yield a measurable sample. 'Measurable sample' and 'measurability' are not precise or formal concepts, but are used to identify a set of practical criteria which can be useful in distinguishing from others those sample designs and procedures "which allow the computation, from the sample itself, of valid estimates or approximations of its sampling variability" (Kish 1965, Section 1.6; also see discussion in the Survey Statistician Nos. 13-15, 1985-86; and Kish 1987, Section 7.1).

[1] Firstly, for measurability it is highly desirable that the sample be a probability sample, ie be based on selection procedures which assign known and non-zero probabilities to all elements in the population. Certainly this precludes judgement or purposive samples, non-probability selections for experimental design, arbitrary selections of single sites for 'case studies', and most quota samples. However concerning the last, useful indications of the level of uncertainty due to sampling variability can sometimes be produced from quota samples of sufficiently large size with many control categories and only small quotas taken from any one category. Examples are good opinion or market research surveys which routinely include useful indicators of margins of uncertainty due to sampling. In normal survey practice one also encounters samples which are not probability samples in the strict sense, in that only relative rather than actual selection probabilities are known. In such samples it is possible to produce valid estimates of proportions, means and ratios, etc, as well as estimates of their variances; but this may not be the case concerning estimates of

population aggregates without the importation of external information. In this a sample may be measurable with respect to certain statistics but not with respect to others. Hence what is required for measurability is: (i) preferably, that the design is such that the actual probabilities, or at least the conditional probabilities of inclusion in the sample given the initial sample size, are positive and known; or (ii) as useful extension in some cases, that any non-random procedures involved are controlled and the population to which they are applied can be considered reasonably randomised. The second conditions cannot be usually ensured in national household surveys, and it is necessary to rely primarily on (i) for measurability.

[2] The second requirement is that of having two or more internal replications in the sample for each domain or stratum for which separate estimates of variance are required. Estimates of sampling variance (and of other components of variable error) can be produced from the sample results themselves only from comparisons between estimates from parts of the sample assumed independent. Most simply this means that two or more independent selections must be available from each stratum. Most practical designs do not meet this requirement exactly, but in many situations it is approximated closely enough for valid estimates of variance to be produced. Examples of designs which can be considered measurable in this sense are many systematic samples, samples with single selections per stratum, 'controlled selection' in which sampling across strata is linked in some manner, and many situations in which the multiple selections within strata are not fully independent. Several of these features are introduced into practical designs for efficiency and/or convenience.

[3] In practice it is necessary to go further then the above requirement of a minimum of two replications. It is necessary to produce valid estimates of the statistics and their standard errors. By 'valid' we mean that the variance estimator generated is sufficiently accurate to be useful, ie is not subject to unacceptably large variances and bias. This is also a practical criterion: the required accuracy depends on the use to be made of the information on sampling errors. Clearly, this information is secondary to the main substantive findings of the survey, and for many purposes it is sufficient to obtain only approximate values of the sampling errors. Furthermore, sampling errors represent just one component of the total error and not in all circumstances the most important one. Nevertheless, in large-scale national household surveys, reasonably accurate estimates of sampling errors for many variables and subclasses are essential for proper interpretation of the results and for evaluation and improvement of sample design.

At a minimum, 'validity' requires that in computing variances the actual sample design and estimation procedures are taken into account. Furthermore, the sample should be large enough in size (in terms of primary selections, replications, or other relevant computing units) to yield reasonably stable estimates of variance. In the sample design it is also necessary to avoid extreme variations in unit sizes and selection probabilities.

[4] Fourthly, the procedures for estimating each statistic of interest and its sampling error need to be formulated in accordance with the sample design. In most cases this is not a major problem because several good general variance estimation procedures applicable to complex statistics in complex designs have been developed as described in Chapters 2 and 3. Nevertheless there can be situations for which valid procedures cannot be formulated on the basis of existing methodology. A requirement of good survey practice is to avoid getting into such situations.

[5] Finally, it is necessary to have the means to implement appropriate procedures of variance estimation in practice. It is necessary to have suitable computer software and full documentation of the structure of the sample for this purpose. It is highly recommended that sample weights, PSU and stratum identifiers, and other essential information on the sample structure be included as an integral part of the computerised micro-level data files resulting from the survey. Failure to ensure such information can render 'unmeasurable' samples which are measurable in other respects. There are many examples where the survey documentation does not meet this basic requirement, even in developed countries, as for instance Kish et al (1976) found in their attempts to compute sampling errors from archived survey data in the United States. Kish notes that "even today variance computations are not feasible for most multistage probability samples, because the needed identification of strata and primary selection number are lacking from computer tapes" (Survey Statistician, No 13). The author also cites the positive example of the World Fertility Survey in developing countries in contrast to much survey practice elsewhere: "All the WFS samples from less developed countries have and will have measurability: their sampling errors can and are being computed. But none of the parallel fertility surveys from the developed countries of Europe have yet computed sampling errors, and perhaps cannot do so now." (Verma et al 1980, discussion).

4.2 SELECTING STATISTICS FOR VARIANCE COMPUTATION IN MULTISUBJECT SURVEYS

4.2.1 GENERAL CONSIDERATIONS

Diversity of Statistics

National household surveys are typically large-scale and multipurpose, and involve the production of separate estimates for a very large number of statistics; the analysis of the results in the form of detailed tabulations, for instance, can involve thousands or even tens of thousands of cells. The great multiplicity of estimates arises from several sources. (1) Most surveys involve the collection of information on a number of substantive variables. In practice there are hardly ever any true uni-subject surveys. (2) For the same set of variables it is usually required to compute many different types of statistics ranging from aggregates and proportions to indices, differences and other functions of ratios, and more complex measures of distribution and relationship. (3) Separate estimates in more or less full detail are often required for geographical and other subnational domains. (4) The greatest increase in the number of statistics perhaps comes from the need to produce separate estimates for diverse subclasses of the population. Meaningful analysis of the survey data usually requires the classification of the survey units in many different ways and in great detail. (5) Also, most surveys involve comparison between subgroups. With many subclasses, the number of subclass differences of interest can become almost unlimited.

Criteria for Choice

In view of the above, it is necessary to be selective in the choice of statistics for which to compute sampling errors. The scope of computations in any particular survey should be decided on the basis of its specific objectives and requirements, taking into account practical constraints.

Sometimes selective computations covering only the most important statistics of interest are all that is possible or required. With related surveys or rounds similar in content and design, it is often unnecessary to compute the full set of sampling errors afresh for every survey.

However, in large-scale national household surveys it is generally quite inappropriate to confine the computations to a few arbitrarily selected statistics, or to seek unnecessary short-cuts and crude approximations, or to rely solely on imputing information from other situations, rather than supplementing any existing information on sampling errors with fresh computations where possible. There are several reasons for recommending as extensive a set as possible of sampling error computations in each survey. (1) In programmes of household surveys, it is very useful to accumulate information on sampling errors covering a wide range of statistics, designs and situations. Such information can help in the design and analysis of future surveys. (2) Results of individual computations are subject to great variability. Appropriate averaging over many computations can yield more stable and useful results. (3) It is necessary to extrapolate or impute information to statistics for which computations have not been made. Such extrapolation requires identification of the patterns of variation of sampling error results as discussed in detail in Chapter 6. The identification of patterns requires many computations covering diverse statistics. (4) The evaluation and improvement of sample design also requires the identification of such patterns.

In the selection of statistics for sampling error computation, the objective should be to capture the widest possible range of values encountered in the survey. The reference here is not so much to the actual magnitudes of standard error, but to certain other parameters or components which may be derived from them and to the general pattern of the results. For instance, one should try to cover statistics with diverse design effects, and a wide range of coefficients of variation and other measures which are useful in identifying the general pattern of variation of the results. This point will become clearer after discussion of 'portable measures' in Chapter 6.

Derived Statistics

One of the important objectives of sampling error computations is to identify the underlying patterns and relationships in the results for diverse statistics. This is greatly helped by computing, in addition to actual standard errors, a range of more 'portable' derived measures such as relative errors, design effects, 'rohs', coefficients of variation and similar measures as explained more fully in Part II. Also useful for the same purpose is information on variation in cluster sizes and the manner of distribution of subclasses of different types over sample areas.

One objective of the derived statistics of the type referred to above is to separate out the effect of various features of the sample design on the magnitude of the sampling error. While analysis of the error into components by sampling stage may not feasible, it is often possible to identify the effects of weighting, cluster sizes, and perhaps also of some important aspects of the estimation procedure.

4.2.2 VARIABLES AND STATISTICS

Substantive Variables

The first priority should be given to covering the widest possible range of substantive variables included in the survey. This is because usually the pattern of sampling errors differs most markedly across substantive variables, compared for example with the variation across subclasses for a given variable. Extrapolations and imputations can often be made more easily across different subclasses or sample bases for a given variable than across different variables. At a minimum, sampling errors should be computed for all important means and proportions with the total sample as the base. The requirement to cover most variables in the survey is usually not difficult to meet because, even in complex multi-subject surveys, the number of important substantive variables is usually not large: a typical survey may involve no more than, say, 30-60 such estimates. The great range of statistics encountered arises primarily from the need to produce separate estimates for numerous subclasses and comparisons.

Often it is possible and useful to group survey variables on the expectation of similarities in the pattern of sampling error results. The grouping may be based on substantive considerations as well as any available information on sampling errors. The objective is to make the groups homogeneous. In selecting the set for sampling error computations, it is important to cover as many different groups as possible, rather than many variables from only a few groups.

Types of Statistics

It is also useful to cover various types of statistics (such as estimates of aggregates, proportions, means and other ratios) because the pattern of sampling error results may differ greatly among them. Of course, for certain types of statistics the pattern may be simpler and more easily related to other statistics; for them more selective computations may suffice. It is desirable to cover more thoroughly statistics with more complex patterns of sampling errors. For instance, it is usually more useful to cover many means and ratios, while proportions, especially those defined in terms of similar characteristics, can be covered more selectively.

4.2.3 DOMAINS AND SUBCLASSES

Geographical Domains

In many surveys, statistics similar to the national level are also produced for a number of urban-rural, regional or other geographical domains. This has a multiplicative effect on the number of statistics for which sampling errors are required. Hence an important question is whether it is necessary and useful to compute the full set of sampling errors for each domain. The answer to this depends on the number of domains involved and on how different they are in terms of the nature of the population covered, survey conditions, sample design, size and nature of the sampling units, and other factors which affect the magnitude and pattern of sampling errors. Such differences are usually marked between urban and rural areas, thus necessitating separate computations. By contrast, conditions are often more similar across different regions of the country, and regions can also be more numerous, especially in large countries. In such circumstances, one may confine the computations at the regional level to only a subset of the most important statistics and explore the extent to which errors computed at the national level may be extrapolated to the regional level. A common difficulty in computing variances at the domain level is that individual domain samples are based on small numbers of primary selections, resulting in unstable variance estimates. This also favours pooling and averaging of individual computations and their extrapolation to other situations in an appropriate manner.

Subclasses

Much more numerous are subclasses defined in terms of characteristics of individual units, hence necessitating a high degree of selectivity in computing sampling errors. In selecting subclasses, priority naturally has to be given to those which are the most important in substantive tabulation and analysis of the survey results. For instance, in demographic surveys most analyses involve classification of the samples by age, sex and other demographic characteristics of individuals: sampling errors for subclasses defined in terms of these characteristics are therefore important. Similarly, in income and expenditure surveys, classifications by household size and composition may be the most important.

The pattern of results over subclasses depends on the substantive characteristic defining the subclass, how the subclass is distributed over the population and sample clusters, and the size of the subclass. It is desirable to cover subclasses of different types and sizes. It is more useful to cover subclasses defined in terms of different characteristics than to cover many categories of the same characteristic. Similarly, subclasses may be grouped according to the manner in which they are distributed over sample clusters, and it is desirable to cover some subclasses from each group. In this context it is useful to note a classification proposed by Kish et al (1976).

[1] The basic concept is that of a cross-class, the members (elements) of which are more or less uniformly distributed across the sample areas and strata. Examples are many subclasses defined in terms of individual characteristics of households or persons such as sex and age groups, which tend to be well distributed across the whole population and hence across sample clusters.

[2] At the opposite end, we have geographical or segregated classes which are largely confined to only a subset of the sample areas. These are similar to but not the same as what we have termed geographical domains in that the latter are explicitly defined as area domains.

[3] As an intermediate category, we have mixed classes which are neither highly segregated nor well distributed like cross-classes. Examples are certain classes defined in terms of socioeconomic characteristics, such as persons with higher education or in certain occupations which are highly unevenly distributed between urban and rural areas.

While technical details are discussed in Chapter 6, some important features of subclass sampling errors should me noted here. Empirical results show that the pattern of variation of sampling error with subclass size is closely related to how the subclass elements are distributed over the population. Sampling errors increase with decreasing subclass size: in a simple random sample in inverse proportion to the square-root of the subclass size, but not quite proportionately in complex samples. This is because generally, design effects (defts) decline as we move from the total sample to estimates over subclasses. In well distributed cross-classes the decline in deft with decreasing subclass size tends to be more marked than in less well distributed classes; consequently in the former the increase in sampling error with decreasing sample size tends to be less steep. In any case, the pattern of variation of the subclass sampling errors depends on the nature of their distribution in the population.

Another factor, though generally less important than the above, is the correlation between the characteristic defining a subclass and the substantive variable being estimated. Subclasses defined in terms of characteristics which are closely related to the substantive variable being estimated often tend to be more homogeneous than the population at large.

This tends to reduce the sampling error for the subclass categories concerned, or moderate the increase as we move from the total sample to subclasses.

In situations where the total sample design effects themselves are close to 1.0 (ie where there is little impact of clustering of the sample), the need to compute sampling errors over many subclasses is reduced. It is more important to study the patterns of variation across subclasses of different types and sizes for heavily clustered designs with large design effects, and especially for variables with the largest defts. In any case, the appropriate strategy is to begin by computing sampling errors (including design effects) for diverse statistics with the total sample as the base, and if necessary separately for different geographical domains, and then decide as to what extent and for what types of subclasses additional computations will be most useful.

Subclass Differences

Almost all surveys involve many comparisons across subgroups and/or over time. Hence it is important to have information on estimates of differences between subclasses and between samples over time. Fortunately in many circumstances it is possible to be quite selective in computing sampling errors for differences and comparisons, in so far as the pattern of results can be deduced from the results for individual subclasses. For instance in comparisons between subclasses based on independent samples, the variance of the difference is simply the sum of variances of the classes being compared. This generally applies to comparisons between geographical domains and often also to those between segregated classes.

For mixed or cross-classes, where the groups being compared come from the same primary selections, the variance of the difference needs to take into account the covariance term. However, even here the effects of clustering and stratification often rapidly decline with decreasing subclass size, making the pattern of sampling errors for differences simpler and more predictable, thus reducing the need for many separate computations. In summary, it is more important to compute sampling errors of differences for subclasses which are overlapping and large in size, and for variables with large subclass design effects.

ILLUSTRATION 4A COMPREHENSIVE SETS OF SAMPLING ERROR COMPUTATIONS

The national fertility surveys conducted under the World Fertility Survey (WFS) programme in 42 developing countries provide an example of one of the most systematic efforts in computing and presenting sampling errors for a wide range of statistics from a complex survey. The example below illustrates the points discussed above concerning the selection of statistics for variance computation. The core of the surveys involved interviewing women in the child-bearing ages on their demographic and background characteristics, marriage and birth histories, knowledge and use of contraceptive methods, and preferences concerning child bearing.

Table 4A.(1) shows some results on sampling errors from the WFS. Countries included here are only a small subset of the WFS surveys, for almost all of which comprehensive information on sampling errors is available in the national survey reports. The results shown here are taken from a comparative study (Verma, Scott and O'Muircheartaigh, 1980); the actual sets of variables and subclasses covered in the national surveys differed somewhat from country to country.

The first column in the table shows a set of variables selected for sampling error computation. This set covers most of the substantive topics of interest in the surveys. Note that even in a complex survey, this number of variables is quite manageable. The variables have been grouped according to topic. There is also a certain degree of homogeneity within the groups in terms of the pattern of sampling errors (as indicated from example by the design effects shown). Most variables of interest were in the form of proportions or means, the estimation of population totals or aggregates not being an objective of the surveys. The set shown includes most of the important means; the coverage of proportions is more selective because of similarities in the pattern among different proportions of the same type.

Before considering the computation of sampling errors over diverse subclasses, it is necessary to examine the results for the set of variables computed over the total sample. Rather than the actual values of standard errors, it is more useful to study a derived statistic such as the design effect (deft) which is more directly comparable across diverge domains and subclasses of the sample for a given variable or set of variables. As detailed in Chapters 5 and 6, deft values are often similar across domains of similar design; across well distributed subclasses, defts tend to decline with subclass size. It is more important to compute sampling errors for different domains and diverse subclasses for designs with large defts, especially for variables with particularly high values. Computations for subclasses are less necessary, and hence can be undertaken more selectively, in cases where the total sample defts themselves are small (close to 1.0).

The total sample defts are relatively large in Table 4A.(1) for countries like Nepal, Mexico, Thailand, Indonesia and Colombia, where the deft values range from 1.5 to 2.3; by contrast, the overall deft values are smaller in the fertility surveys of countries like Bangladesh and Fiji. (Countries have been arranged in the table in order of decreasing deft averaged over all variables over the total sample.) In the WFS surveys in Sri Lanka, Guyana, Jamaica and Costa Rica (not shown here, but available in the original source to the table), defts computed with the total sample as the base averaged under 1.2.

The subclasses of interest are determined according to the tabulation and analysis requirements of the surveys. In the present case, subclasses representing demographic groups (age, marriage duration, family size groups, etc) are the most important because they are closely related to the variables being studied in the surveys. Next come the classes defined in terms of socioeconomic characteristics, which are important in the study of differentials. Table 4A.(2) specified a typical set. Regarding geographical domains, the requirements are much more country-specific, despite the common content of the surveys being considered here. However, the requirement for separate urban and rural computation is common.

Another example of comprehensive work in the area of sampling errors is provided by the equally wide-ranging series of surveys conducted more recently in developing countries under the Demographic and Health Survey (DHS) programme. The content of the surveys parallels the WFS, except for a much wider coverage of variables related to mother and child health but somewhat less elaborate coverage of socio-economic characteristics of the survey respondents. On the following page Table 4A.(3) provides a typical example of variables selected for sampling error computation, and Table 4A.(4) shows the total sample defts, taken from a comparative study of the DHS surveys (Aliaga and Verma, 1991). The selection of results shown here complements that from the WFS by covering a number of mother and child health (MCH) variables, and also by providing data from seven African countries. Together, the two series provide perhaps the most impressive set of comparative data on sampling errors from surveys in developing countries, albeit in a particular subject matter area. Frequent reference will be made to these results in the following chapters to illustrate various practical issues in the computation and analysis of survey sampling errors.

TABLE 4A.(1). List of variables and defts based on the total sample from the world fertility survey.
(Source: Verma, Scott and O'Muircheartaigh, 1980)

Design factors (deft) for the total sample

	Nepal	*Mexico*	*Thailand*	*Indonesia*	*Colombia*	*Peru*	*Bangla-desh*	*Fiji*
NUPTIALITY								
01 %, currently married	1·14	1·68	1·06	1·35	1·26	1·02	1·05	1·07
02 %, exposed to child-bearing	2·11	1·13	1·59	1·36	1·07	1·25	1·06	1·36
03. %, with marriage dissolved		1·41	0·98	1·44	1·47	1·09	1·13	1·21
04 %, remarried		1·12	1·49	1·51	1·24	1·12	1·39	1·23
05 number of marriages	–	1·36	1·30	1·79	1·69	1·10	1·41	1·47
06. age at marriage	2·49	1·76	1·28	1·54	1·35	1·05	1·48	1·65
07. time spent within marriage	–	1·25		1·14	1·09	1·04	1·04	0·97
FERTILITY								
08. %, pregnant	1·39	1·06	1·02	1·31	0·86	1·26	1·10	1·30
09 children ever-born	2·08	1·78	1·52	1·34	1·15	1·07	1·05	0·98
10. living children	1·96	1·78	1·29	1·43	1·11	1·12	1·00	0·94
11. births in first 5 years	3·02	1·44	0·98	1·38	0·88	0·87	1·24	1·05
12. births during past 5 years	2·74	1·58	-	1·54	1·31	1·04	1·03	1·00
13 first birth interval	1·38	1·22	1·16	1·44	1·25	1·14	1·20	
14 last closed birth interval	1·41	1·32	1·43	1·43	1·04	1·02	1·17	0·89
15 open birth interval	2·05	1·61	1·37	1·56	1·14	0·92	1·20	
16 months breast fed child	2·08	2·05	1·98	1·46	1·52	1·50	1·12	
17. 16, excluding dead children	1·26	1·45	1·75	1·28	1·50	1·09	1·16	
18 %, of children who died	1·70	2·08	1·30	1·34	1·63	1·81	1·08	1·11
FERTILITY PREFERENCES								
19. %, last pregnancies unwanted	–			1·44	1·28	1·34	1·10	1·63
20 %, wanting no more children	2·39	1·36	1·29	1·71	1·12	1·18	1·28	1·62
21. %, expressing boy-preference	2·13	1·06	1·15	1·34	1·06	0·97	1·23	0·94
22 %, exceeded desired family size	2·37	1·49	1·15	1·31	1·46	1·21	1·17	1·29
23 additional children wanted	3·16	1·96	1·62	1·62	1·09	1·42	1·23	1·52
24. desired family size	3·76	1·74	1·63	1·86	1·18	1·32	1·26	1·22
CONTRACEPTIVE KNOWLEDGE								
25. %, knowing pill	3·25	3·20	2·74	2·32	2·84	1·74	1·70	
26 %, knowing IUD	2·95			2·58	2·30		1·71	-
27 %, knowing condom	2·44	2·50	1·94	2·60	2·38	1·74	1·42	1·66
28 %, knowing modern method	4·19	2·74	2·77	2·24	2·46	2·10	1·80	
CONTRACEPTIVE USE								
29. %, ever used pill		1·90	2·05	1·95	1·82	1·26	1·30	1·32
30 %, ever used IUD				1·90	1·21		1·05	
31 %, ever used condom				1·56	1·13		1·52	
32. %, ever used any method	2·23	2·19	2·28	1·79	2·35	1·84	1·56	1·42
33. %, ever used modern method		2·08	2·34	1·82	1·93	1·33	1·54	1·44
34 %, used in open interval		1·96	1·78	1·70	2·05	1·46	1·38	
35 %, used in closed interval				1·30	1·69	1·47	0·99	1·26
36 %, currently using any method		1·96	2·08	1·67	1·98	1·40	1·36	1·40
37 %, currently using modern method		1·83	2·35	1·72	1·49	1·06	1·22	1·29
38. 37, for women wanting no more children		1·50	2·19	1·52	1·34	1·06	1·20	1·02
SAMPLE SIZE (ever-married women)	5940	6255	3820	9136	3302	5640	6513	4928
No. of effective PSUs	40	152	70	376	405	410	240	100

TABLE 4A.(2). Example of the subclasses used in 4A.(1).

Definition of Subclasses Used

The term *subclass* is used to refer to a subset of the sample defined in terms of a particular attribute of the individuals in the sample. Sampling error of each variable was computed for the following sets of subclasses in most countries. Note that only the demographic subclasses are strictly comparable between countries. To compute subclass differences the subclasses were taken in pairs in the order in which they have been listed below.

Demographic subclasses

Age : Women aged under 25 (M_s = 0·25);† women aged 25–34 (M_s = 0·35); women aged 35 44 (M_s = 0·27); women aged 45 49 (M_s = 0·13).
Marriage duration : First married less than 5 years ago (M_s = 0·20); 5–9 years ago (M_s = 0·20); 10–19 years ago (M_s = 0·32); 20 or more years ago (M_s = 0·28).
Children ever born : 3 or fewer children (M_s = 0·50); 4 or more children (M_s = 0·50).

Socio-economic subclasses

Age at first marriage : under 25; 25 and over (under 20, and 20+ in Bangladesh and Nepal).
Woman's literacy (1) : not literate; literate.
Women's literacy (2) : as above, but confined to women currently aged 25–34.
Husband's level of education : no schooling; attended primary but not completed; completed primary; secondary or higher.
Husband's occupation : technical, administrative or clerical; sales and services; skilled or unskilled manual; farming.
Religion, ethnic group, etc. : as relevant.

Segregated classes (geographic domains)

Type of place of residence : urban; rural.
Region : usually 3–6 major regions of the country, as relevant.

† M_s = approximate size of the subclass as a proportion of the total sample of women

TABLE 4A.(3). List of variables for sampling error computations for DHS (Source: Aliaga and Verma, 1991)

group		code	type*	description
fertility and related				
1	1	BBEFXX	p	birth before a certain specified age
1	2	CDEAD	r,m	proportion (or mean number) of children dead
1	3	CEB	m	mean number of children ever-born
1	4	CEB40	m	completed fertility (to women aged 40+)
1	5	CMAR	p	proportion currently married
1	6	CSUR	r,m	proportion (or mean) of children surviving
1	7	EXPOS	p	proportion exposed
1	8	PREG	p	proportion pregnant
1	9	SINGLE	p	proportion single
health				
2	10	ATTE	r	proportion of deliveries attended
2	11	BCG	r	proportion of children receiving BCG
2	12	COUGH	r	proportion of children having cough
2	13	DIAR	r	proportion of children having diarrhoea
2	14	DIATR	r	prop. of children receiving treatment for diarrhoea
2	15	DPT	r	proportion of children receiving DPT
2	16	FEVER	r	proportion of children having fever
2	17	FULLIM	r	proportion of children receiving full immunisation
2	18	HCARD	r	proportion of children having health card
2	19	HTAGE	r	height for age
2	20	MEASLE	r	proportion of children having measles
2	21	POLIO	r	proportion of children receiving polio imm.
2	22	TETA	r	proportion of children receiving tetanus imm.
2	23	TREATC	r	proportion of children receiving treatment for cough
2	24	TREATF	r	proportion of children receiving treatment for fever
2	25	WTAGE	r	weight for age
2	26	WTHGT	r	weight for height
fertility preferences				
3	27	DELAY	p	proportion wanting to delay next birth
3	28	IDEAL	m	mean ideal family size
3	29	NOMORE	p	proportion wanting no more children
contraceptive knowledge and use				
4	30	CUSE	p	proportion currently using contraception
4	31	EVUSE	p	proportion ever-used contraception
4	32	KANY	p	proportion who know of any method
4	33	KMOD	p	proportion who know of any modern method
4	34	KSOURC	p	proportion who know of any source of method
4	35	UCOND	p	proportion ever used condom
4	36	UIUD	p	proportion ever used IUD
4	37	UMOD	p	proportion ever used any modern method
4	38	UPIL	p	proportion ever used the pill
4	39	USTER	p	proportion ssterilised
proximate factors				
5	40	ABST	m	mean length of post-partum abstinence
5	41	AMEN	m	mean length of post-partum amenorrhoea
5	42	BF	m	mean length of breast feeding
5	43	UABST	p	proportion ever used abstinence
5	44	UTRAD	p	proportion ever used traditional method
background characteristics				
6	45	EDUC	p	proportion with higher education
6	46	MBEFXX	p	proportion married before a certain age
6	47	NOED	p	proportion with no education

* p=proportion of women; m=mean per woman;
r=ratio of two substantive variables.

TABLE 4A.(4). Computed design effects (defts) for the total sample - Demographic and Health Surveys (DHS) programme. (Source: Aliaga and Verma, 1991).

group	s.no.	variable	GHANA	ZIMBABWE	UGANDA	BOTSWANA	DOMINICAN REP.	PERU	BRAZIL	EGYPT	ECUADOR	SENEGAL	KENYA	THAILAND
			10	20	30	40	50	60	70	80	90	100	110	120
1	1	BEFX		1.26	1.47	1.16							1.34	
1	2	CDEAD	1.24	1.24		1.29	1.14	1.20	1.14		1.28	1.33		
1	3	CEB..	1.02	0.99	1.03	1.33	1.33	1.09	1.31	1.46	1.29	1.15	1.51	1.84
1	4	CEB40	0.98	0.98	1.18	1.02	1.16	1.21	1.12	1.57		1.18	1.34	1.59
1	5	CMAR.	1.50	1.06	1.30	1.53	1.43	1.12	1.43		1.38	1.87	1.59	
1	6	CSUR.	0.99	0.94	1.02	1.34	1.30	1.05	1.28	1.37	1.26	1.10	1.48	1.73
1	7	EXPOS					1.22	1.09	1.34		1.15	1.16		
1	8	PREG.	1.19	0.79	1.12	1.09	1.09	1.05	1.12	1.20	1.04	0.97	1.54	
1	9	SINGL					1.60	1.16	1.35		1.41	1.77		
2	10	ATTE.	1.21	1.38	1.23	2.11	1.63	1.35		2.49	1.68	2.03	2.10	2.17
2	11	BCG..			1.01		1.28	1.01		1.62		1.17	1.23	
2	12	COUGH	1.33	1.25		1.66							1.39	
2	13	DIAR.	1.35	1.18	1.34	1.61	1.33	1.08		1.14		1.40	1.07	1.48
2	14	DIATR	1.50	0.98		1.25	1.23	1.13		1.23		1.03	1.07	1.17
2	15	DPT..			1.18		1.30	1.06		1.42		0.87	1.52	
2	16	FEVER	1.39			1.79							1.23	
2	17	FULLI	1.20	1.05	1.24	1.44							1.24	
2	18	HCARD	1.43	1.20	1.26	1.18	1.35	1.07		0.99			1.01	1.63
2	19	HTAGE	1.08	1.30	1.07									
2	20	MEASL			1.42		1.33	1.04		1.54		1.00	1.33	
2	21	POLIO			1.19		1.11	1.05		1.34		1.11	1.26	
2	22	TETA.	1.91	1.30	1.63	1.70	1.28	1.28		1.83	1.60		1.42	1.83
2	23	TREATC	1.18	1.18		1.14							1.54	
2	24	TREATF	1.58			1.26							1.47	
2	25	WTAGE	1.04	1.33	1.19									
2	26	WTHGT	1.05	0.94	1.06									
3	27	DELAY	1.21	1.10	1.08	1.35	1.19	1.08	1.26	1.11	1.08	1.13	1.30	1.41
3	28	IDEAL	2.14	1.33		1.55	1.20	1.43	1.38	1.64	1.21	1.64	2.11	2.35
3	29	NOMORE	1.36	1.25	1.20	1.14	1.26	1.09	1.29	1.50	0.95	1.26	1.35	1.37
4	30	CUSE.	1.16	1.14	1.17	1.26	1.31	1.21	1.22	2.27	1.42	1.35	1.49	1.54
4	31	EVUSE	1.76	1.22	1.66	1.51	1.35	1.38	1.31	2.70	1.62	1.82	1.70	1.72
4	32	KANY.	2.48		1.48	2.61		1.53			2.19	1.54	2.37	
4	33	KMOD.	2.57	1.26	1.66	2.67		1.58			2.18	1.73	2.66	
4	34	KSOURCE			1.61				1.24	1.66			2.51	
4	35	UCOND								1.32				
4	36	UIUD.								1.89				2.10
4	37	UMOD.	1.21	1.20		1.27							1.29	
4	38	UPIL.					1.19	1.19	1.20	1.66	1.28			1.86
4	39	USTER					1.26	1.15	1.22		1.24			1.88
5	40	ABST.	1.40		1.34	1.41	1.04	1.02	1.20	1.20	1.09	1.40	1.42	1.25
5	41	AMEN.	1.32		1.09	1.24	1.05	1.09	0.97	1.24	1.15	1.01	1.37	1.36
5	42	BFEED	1.21	1.09	1.19	1.12	1.14	1.09	1.15	1.24	1.11	0.85	1.12	1.51
5	43	UABST						1.07	1.28		1.13			
5	44	UTRAD			1.05		1.16	1.09			1.23	1.36		
6	45	EDUC.	1.87	1.73	1.82	1.77	1.98	1.53	2.28	3.07	1.95	2.02	2.33	2.10
6	46	MBEFX	1.26	1.43	1.57	1.50	1.73	1.26		2.24	1.43	1.82	1.73	
6	47	NOED.								2.96		2.23	1.98	

4.3 CHOICE OF THE METHOD

The strengths and limitations of the various practical methods for computing sampling errors for large-scale surveys have been indicated in the description of the methods in Chapters 2 and 3. The most important considerations are summarised below with a view to indicating the method or methods which might be the most suitable choice in particular circumstances.

Criteria in the Choice of a Method

Several criteria are involved in the choice of a method in practice. Roughly in order of importance, these include:

[1] availability and convenience of software for application of the method;

[2] issues relating to computational convenience, economy and speed;

[3] the type and range of statistics for which sampling errors are required;

[4] the complexity of the estimation procedures involved;

[5] how well the sample structure fits or approximates the model assumed for application of the method;

[6] statistical properties of the various methods; and

[7] suitability of the method to meet special requirements such as analysis of variance components or estimation of non-sampling components of variance.

Linearisation

The linearisation approach has certain advantages in being able to handle a diversity of designs more easily, especially for the estimation of means and proportions. This can be important in the context of household survey programmes where a number of surveys with different designs may be involved. Another advantage is that less computational work is involved than the alternative approaches based on repeated replication. Despite the rapid improvement in computer facilities, the volume of computational work can still be an important consideration in the choice of a particular method, especially in developing countries. In fact, the difference between the methods in this respect becomes more marked for samples with many strata and PSUs, because with repeated replication, the number as well as the average size of replications increases in proportion to the number of sample PSUs. it is an important point because it is not uncommon in developing country surveys to use samples with several hundred PSUs (Illustration 3A).

However, perhaps the critical factor favouring this method is its advantage as regards the availability of general purpose 'portable' software for its application. Many developing country organisations are not in a position to develop and maintain their own special purpose software. (Section 4.6.)

Its main limitation is the difficulty (or at least the need for special procedures) in dealing with complex statistics and estimation procedures.

Repeated Replication

Repeated replication methods also have their advantages, and represent the only available choice for certain special purposes. They are especially appropriate for handling complex statistics (such as coefficients in multiple regression); and even more importantly, for complex estimation procedures involving a number of adjustments to the data - which are not so uncommon in survey analysis even in developing countries. Some examples of such application are given in Illustration 5B.

Of the two main repeated replication methods, JRR is preferable for being technically the simpler and also somewhat more flexible, though generally it involves more computational work than BRR. The BRR method is technically more complex, and also more restrictive in the sample designs handled, or at least the extra steps required in dealing with designs other than the basic paired selection model. In designs with many small PSUs, it may be necessary to define more suitable - generally larger, fewer and more uniform computing units. This requires care in ensuring that the redefined units still properly reflect the actual sample structure: otherwise the variance estimation may be biased. An example of such redefinition is provided in Illustration 4B below.

Nevertheless, while the linearisation approach remains the most commonly used method, the use of repeated replication procedures is increasing, especially for complex statistics.

The Overriding Consideration.
Regarding the comparison between linearisation, JRR and BRR, the main conclusion is that in many circumstances any of these methods can provide satisfactory results: when judged by several criteria none of them has been found to be strongly and consistently better or worse (Kish 1989, 13.5). Therefore the choice among them is chiefly determined by practical considerations of convenience, cost and availability of software.

Independent Replications

Note should also be taken of the possible uses of the simple independent replication approach in certain circumstances, as noted in Section 3.2. Where the sample design permits, there can be advantages in considering the simple replicated approach, especially in developing country circumstances. This is despite the various (and some serious) limitations of the method. The simplicity of the approach means that, where applicable, the method will encourage the computation and presentation of sampling errors. This is important because at present, the common situation unfortunately is that in many surveys no information whatever is provided on sampling errors. With the simple replicated approach no special-purpose software is required and some information on sampling errors can be generated simply as a by-product of the normal process of tabulation of the survey data. The computational cost is the need to produce separate estimates for each replication.

It is also not necessary to rely on the independent replication approach as the sole or even the main method of variance estimation: where applicable, the method can be used to complement the results from more precise methods of variance estimation. It can be used to obtain quickly and economically rough estimates of sampling error for a wide range of statistics, from which general patterns may be identified and subsets for more precise computations using more sophisticated methods selected. The simple method can also be used to check the results of more complicated approaches for gross computing errors. Actually, comparisons among replicated estimates can be helpful more generally in identifying unusual patterns in the survey results which indicate the need for rechecking the survey procedures.

4.4 FITTING THE SAMPLE STRUCTURE TO THE ASSUMED MODEL

Variance estimation requires two or more replicates selected independently from each stratum. As already noted in Chapters 2 and 3, in practical applications it is sometimes necessary to combine or redefine actual PS's and strata to obtain the primary units and strata to be used in the computations. This may be necessary for several reasons.

(1) It is common in practice that the actual sample structure does not exactly correspond to the model required in the variance estimation procedure. To apply the procedure some additional assumptions have to be made regarding the sample structure. Ideally one would like to ensure that any bias resulting from additional assumptions is unimportant, or at least that it results in 'safe' (conservative or over-) estimates of variance.

(2) Redefining computing units can reduce and simplify computational work.

(3) Redefinition may also be introduced to improve statistical properties of the variance estimates generated, such as reducing their variance or bias, or making their sampling distribution more nearly normal.

4.4.1 COLLAPSED STRATA TECHNIQUE

A common feature of many sample designs is the selection of a single PSU per stratum. This is achieved either explicitly through fine stratification or implicitly through systematic selection. In some designs, special techniques such as 'controlled selection' are used in which more strata (controls) are introduced than the number of primary units which can be selected. An example of the latter is provided by a recent food consumption survey in Indonesia where, because of the complex and intensive nature of the survey, the sample had to be restricted to a relatively small number of PSUs, but it was necessary to simultaneously control for diverse urban-rural, regional, ecological and human factors affecting patterns of food consumption. The use of such techniques helps to improve sampling efficiency, but in general does not permit unbiased estimation of variance. The usual method of variance estimation in such situations is called the collapsed strata technique. Essentially, it amounts to treating each set of two or more PS's as random selections from a single stratum, thus disregarding some of the original stratification. Generally this involves some over-estimation of variance. The routine practice is to minimise the amount of collapsing, only to the point that each newly defined stratum contains the minimum number (namely 2) of PS's essential for variance estimation. Collapsing in pairs also has the advantage of simple variance estimation formulae, which is useful with linearisation methods, but much more important in methods like BRR. (Illustrations 2A and 4B.)

Greater degrees of collapsing, for example in triplets instead of pairs of strata, would tend to increase the overestimation but may also reduce the variability of the computed variances (Kish and Hess 1959; see also 4.4.5 below). Hence in terms of overall accuracy (mean squared error) the best strategy for collapsing strata is a more complicated issue. In a discussion of the issue, Rust and Kaltori (1987) note that while collapsing in pairs "is generally appropriate for national estimates from large-scale surveys with 60 or more PSUs, a greater degree of collapsing may be appropriate when a small sample of PSUs is selected, and especially so when the number of PSUs is as low as, say, 20." The authors also note that a greater degree of collapsing is indicated when the sample size per PSU is small, or when subclass as distinct from total sample estimates are important, or when the differences between the strata means are small.

Since the actual gain from stratification depends on the difference between stratum means, the overestimation due to collapsing strata also depends on the same factor. On the basis of a simple model it can be shown that the overestimation is twice as large as the actual reduction in variance as we move from a 2-PSU per stratum design to a finer stratification with 1 PSU per stratum. In other words, there is an apparent loss in precision equal in magnitude to the actual gain in precision (Cochran 1963, Sec 5A.11). This has an important consequence in practice: collapsing of strata should be done so as to minimise the difference between strata means. In systematic samples this is generally ensured by collapsing adjacent implicit strata, in so far as ordering of the units ensures that similar units come together. In collapsing explicit strata, it is necessary to do so on the basis of stratum characteristics. The important thing to ensure is that this is done on the basis of characteristics of the strata known before the survey, and not on the basis of the characteristics of the particular units which happen to be selected; otherwise the variance may be seriously underestimated. This also requires that information on stratum characteristics is recorded and preserved. (See the warning in Illustration 2B.(2).)

The above also makes it desirable that collapsed strata are similar in size; if not, the variance formulae will need adjustment. Essentially the adjustment is to weight the value for each stratum in proportion to the stratum size. Specifically, if y_1 and y_2 are the estimates for two strata of size S_1 and S_2 respectively, and y the total for the pair, then the expressions $(y_1-y/2)^2$ *and* $(y_2-y/2)^2$ in the paired selection variance formula (equation 2.7, expressed in slightly different notation):

$$var(y) = 2.\left[(y_1-\frac{y}{2})^2 + (y_2-\frac{y}{2})^2\right] \qquad (4.1)$$

are replaced by, respectively

$$(y_1-W_1.y)^2 \quad and \quad (y_2-W_2.y)^2$$
$$where \quad W_1 = \frac{S_1}{S_1+S_2}; \quad W_2 = \frac{S_2}{S_1+S_2}. \qquad (4.2)$$

For ratio means the adjustment is likely to have small effects for moderate differences in the size of the collapsed strata. However a gross overestimation can result in the case of sample totals or aggregates (Kish 1965; Section 8.6B).

In grouping systematically selected PSUs to define 'implicit' or 'collapsed' strata, one should not cut across boundaries of explicit strata within which the systematic selections were conducted independently. (For instance, if an explicit stratum contains an odd number of selections, one of the collapsed strata may be assigned three rather than two units so as not to cut across explicit strata used in sample selection.)

4.4.2 RANDOM GROUPING OF UNITS WITHIN STRATA

The above discussion concerned collapsing of strata leaving the primary units unchanged. It is also possible to group or combine the units in various ways. One technique is to randomly group units within a stratum so as to form larger, fewer, more uniform in size or otherwise more convenient 'pseudo primary selections' to serve as the computing units for estimating variance. This technique is useful when, in the whole sample or some of its strata, the original PS's are too small and/or too numerous for the purpose. The technique is particularly useful in dealing with small subclasses. It is often unavoidable in the application of repeated replication methods.

With random grouping, no additional bias is introduced in the variance estimation; however because of the reduced number of computing units its variability may be increased. On the other hand, there are also some advantages. Grouped units can be made more uniform in size. The technique also reduces computational time, and the distribution of characteristics of the grouped units, based on sums of variables, better approximates normality.

In random samples of units, the appropriate method to group units is to select random subsamples of those units. In a systematic sample, the groups should be formed by selecting 'interlocking' systematic subsamples; for instance if eight units numbered 1 to 8 in a systematic sample are to be grouped into two pseudo units, unit numbers (1, 3, 5 and 7) may appear in one group and the remaining in the other (see Illustration 4B). Grouping of adjacent units (which is sometimes done for convenience) would tend to under represent the benefit of systematic sampling.

4.4.3 COMBINING UNITS ACROSS STRATA

In samples consisting of many small PS's selected independently from many strata, units may be randomly grouped across strata so as to give fewer strata and fewer and larger units for variance computation. As with random grouping of units within strata, the technique does not introduce bias but may increase variability of the variance estimator; it has similar advantages of reduction in computing time and of improved approximation to normality. Combining across strata may be accompanied (preceded) by random grouping of units within strata. A convenient way to apply this method would be to first group PS's within strata as necessary so that each stratum contains the same number of grouped units, and then form the final computing units by taking one such grouped unit from each stratum at random. Deming (1960) terms this technique "thickening the zones".

The creation of independent replications as discussed in Section 3.2, each containing one or more units from <u>every</u> stratum, can be regarded as an extreme example of this procedure.

4.4.4 SO-CALLED 'SELF-REPRESENTING PSUs'

The term 'self-representing PSUs' is sometimes used to refer to units which appear in the sample with certainty. This happens when some of the units described as 'PSUs' in the frame are considered so large and important that they are automatically included into the sample, while elsewhere other units of the same description form the true PSUs and are subject to the sampling process. The description of the former as 'self-representing PSUs', though common and possibly convenient for descriptive purposes, can be confusing and should be avoided. Each such unit actually forms a stratum, and the next stage units within it actually subject to sampling are the PSUs to be used (or appropriately combined as explained above) for computing sampling errors.

4.4.5 VARIABILITY OF THE VARIANCE ESTIMATES

It is important to realise that variance estimates from samples are themselves subject to variability, particularly for samples based on a small number of primary selections. The precision of variance estimation is a complex subject. For reasonably large samples with good control to eliminate extremes in 'cluster sizes' (ie in sizes of the primary selections), a useful approximation to the coefficient of variation of a variance estimator is given by Kish (1965; Section 8.6D) as:

$$cv^2 = \frac{2}{df} \tag{4.3}$$

where "df" is the degrees of freedom, approximately equalling the number of PSUs selected less the number of strata. With 2 PSUs per stratum from H strata, we have df = 2H - H = H. In general, with a total of a primary selections from H strata, we have df = a - H, since one df is lost in taking the squared differences from the stratum mean in each stratum for the estimation of variance. For example, with 48 PS's with 3 from each of 16 strata, we have df = 48-16 = 32; however, with the 48 selections coming in pairs from 24 strata, we have df = 48-24 = 24. Hence with less detailed stratification, the variance estimation is more precise, though the magnitude of the variance itself is generally increased. In so far as the latter is more important, stratification is usually carried out to the maximum extent possible, often even beyond the level of 2 PSUs per stratum - resorting to 'collapsing' for the purpose of variance estimation.

In the case of a systematic sample, it was noted above that the common practice is also to collapse pairs of adjacent strata to form new computing strata for variance estimation. With a greater degree of collapsing than the minimum required (eg. triplets in place of pairs), degrees of freedom and hence precision of the variance estimation is increased; but at the same time the overestimation bias is also increased. An alternative procedure with systematic selection, referred to in Section 2.4 (equation 2.22) is as follows. In place of forming (a/2) non-overlapping pairs of adjacent units, one may utilise all possible (a-1) successive differences among the ordered list of (a) units selected systematically. The advantage of this procedure is that it does not increase the overestimation due to collapsing beyond that with the normal pairing, but it reduces the variance of the variance estimator.

The lack of precision of sampling errors computed for individual geographical domains can be a particularly serious problem because each such domain may contain only a small number of primary selections. What can be done to reduce this problem? The basic 'remedy' is to pool computations from several samples, and replace the results of individual computations by appropriately averaged values, as discussed more fully in Chapter 6. In addition, one may avoid computations for separate domains and impute values from the total sample, perhaps with some appropriate adjustment if necessary. Certain sample designs involve the selection of a small number of large PSUs, but the selection of many small clusters or area units at the second stage. In such cases, an option may be to compute variances by treating the more numerous SSUs as if they were the primary selections, and then to adjust the results for the excluded contribution of the first stage (Kish 1989, Section 14.4). This is useful if the first stage contribution is small or can be estimated from some other source.

4.4.6 CODING OF THE SAMPLE STRUCTURE

To compute sampling errors it is essential that all necessary information on the sample structure is available, ideally on the computer data files, as emphasised earlier (Section 4.1). At a minimum the information should include

[1] Identification of the strata as used in the computations, taking into account any collapsing or other modifications which may have been made to the original stratification, and the procedures adopted to deal with systematic sampling.

[2] Identification of the primary computing units, taking into account grouping or combining of the original selections, and ensuring that at least two such units are available from each stratum.

[3] Weights assigned to ultimate units, ideally as an integral part of the survey micro-level data.

[4] Sampling rates, to compute the finite population correction if relevant.

[5] Identification of the domains and subclasses for which separate estimates are to be produced.

Additional information will be required if the overall variance is to be decomposed into components according to different stages of sampling or estimation, or other aspects of the sample structure. (Chapter 5.)

4.5 REDUCING COMPUTATIONAL WORK

There are several ways of reducing the work involved in the computation of sampling errors: confining the computations to an appropriately selected subsample only; redefining the sample structure for the purpose; simplification of the variance estimation procedure; and imputing errors computed for a subset of statistics to other statistics.

<u>Computation Over a Subsample</u>

A simple option is to base the computations on a subsample of the full sample. The subsample should of course reflect the structure of the full sample; it is also necessary to correctly establish the relationship between the variances corresponding to the full sample and the subsample. This procedure is useful only when the sample consists of a large number of primary units, so that the loss in precision in variance estimation from only a subsample is acceptable. Also the subsampling rate should be small enough so that the saving in computing time more than compensates for the additional cost and trouble of constructing the subsample. Two obvious examples of where the approach may be considered are: (i) computing sampling errors for large census samples which are often attached to the full census to collect additional information; and (ii) sample surveys based on a large number of elements selected in a single stage. Occasionally the technique may also be considered for clustered designs with numerous small PSUs, such as small clusters of households.

The basic idea is to estimate unit variance from the subsample and then to use it in the variance formula for the full sample in the ordinary way. The concept of unit variance is useful when the variance or a component of variance is inversely proportional to the number of units (see Section 6.3 for further discussion). As a simple example, consider a SRS of size n from which a random subsample of size n' units is selected. In the ordinary expression for variance of a mean (corresponding to the full sample)

$$var(\bar{y}) = \left(\frac{1-f}{n}\right)s^2, \quad where \ f = \frac{n}{N} \tag{4.4}$$

the idea is to estimate unit variance s^2 from the random subsample as

$$s^2 = \frac{1}{n'-1}\left[\sum' y_j^2 - \frac{\left(\sum' y_j\right)^2}{n'}\right] \tag{4.5}$$

where the summation is over n' units in the subsample. Similarly in estimating the variance of a total y from a multistage sample

$$var(y) = (1-f).a\left[\frac{\sum' y_i^2 - \frac{\left(\sum' y_i\right)^2}{a'}}{a'-1}\right] \tag{4.6}$$

the quantity in the square brackets has been estimated from a random subsample of a' primary selections from the a primary selections in the full sample. The same idea can be extended in a straightforward way to a stratified sample, by applying the above expression separately for each stratum, provided at least two primary units appear in the subsample for each stratum. Introducing subsampling within sample PSUs would generally result in much more complex relationships between the required variance for the full sample and the computed variance from the subsample - requiring some additional assumptions or modelling in most situations.

With rapidly improving computer facilities, the introduction of subsampling to reduce computing work is hardly worth considering in normal household survey work, though it may be convenient for some other special purposes as noted earlier.

Grouping of Units and Strata

Several techniques have already been mentioned in the previous section. Random grouping of units and combining across strata can be a much more useful means of reducing computational work than for example the method of subsampling. For certain methods (such as the BRR), some grouping and combining is almost unavoidable if the number of units is large and variable across strata.

Simplifying the Variance Estimation Procedure

Some simplification and approximation is already involved in the various practical procedures for variance estimation described in the previous chapters. What makes these procedures 'practical' is that the simplifications introduced greatly reduce the computational work, but generally with only a minor effect on the accuracy of the variance estimations generated. For instance, with the assumption of independent and with-replacement selection of PS's within strata, variance can be usually estimated with only a small approximation, simply in terms of certain quantities aggregated to the PS level.

Beyond such basic assumptions underlying the various practical methods, it is sometimes also convenient to introduce additional approximations to reduce the computational work involved. Here are some examples. The finite population correction is disregarded (often with good justification), or averaged or assumed uniform within strata even though actually it may be much more complex. The same may be done in relation to sample weights if they do not differ greatly between units. In the application of the BRR method for instance, one may resort to 'partial balancing' if the number of strata is too large. In the application of replicated methods generally, some steps in the estimation procedure (such as the application of post-stratification weights) may be applied only once for the full sample, rather than to each replication separately as required by the method if applied strictly. There are many other specific instances where certain (hopefully unimportant) components of variance are ignored with a view to simplifying the computations involved.

Imputing Sampling Errors

The magnitude and pattern of sampling errors may be related across similar surveys, similar subclasses or similar variables on the basis of empirical results and/or appropriate models. This issue is the topic of Chapter 6. Here we emphasize that the establishment of patterns of similarity is potentially the most effective way of reducing the amount of fresh sampling error computation which needs to be done in any particular survey.

ILLUSTRATION 4B DEFINING COMPUTING UNITS AND STRATA

In many practical situations, the sample structure requires some redefinition in an appropriate way before a method of variance estimation can be applied. For example, though the BRR method is not confined to two primary selections per stratum, that represents the most convenient design for the method. Some redefinition of the sample structure may be necessary or convenient in the application of other methods as well.

The example in Table 4B.(2) is based on a survey in Colombia and shows how the sample structure may be specified for the purpose, applying the various ideas discussed in the preceding section. The design consisted of two distinct domains:

[1] Certain large localities were taken into the sample with certainty, as 'self-representing' units. Each locality in fact formed a separate stratum from which a number of smaller areas (clusters) were selected systematically. These clusters formed the effective PSUs in Domain [1].

[2] The second domain was composed of smaller localities. A systematic sample of localities were selected first, and then a sample of clusters taken from each selected locality.

Thus the type of areas (clusters) which form the effective PSUs in the first domain, formed the second stage units in the second domain. Computing strata and PSUs were defined as follows.

In Domain [1], each 'self-representing' locality (marked with a consecutive set of asterisks in the table) was divided into one or more computing strata, each stratum consisting of a set of consecutive areas (the actual PSUs or clusters in the sample). Thus, for example, Barranquilla was divided into two computing strata (numbered 01 and 02 in the table), the first consisting of clusters 1-14 and the second consisting of clusters 15-30, the clusters being numbered in the order of selection. Here, as elsewhere, a desirable objective was to make computing strata and units reasonably uniform in size. Next, in each stratum so defined, alternative clusters from the ordered list were grouped to define two 'interlocked' computing units (replicates) to be used as primary selections for variance estimation. Thus in each of the above two strata, odd numbered clusters formed the first and the even numbered clusters formed the second replicate. This is the 'combined stratum' technique involving the linking of alternative clusters across implicit strata, resulting from systematic sampling, into sets each of which forms a computing unit. It also involves collapsing pairs of adjacent sets so defined to form a computing stratum with two computing units.

The large metropolitan area, from which 136 clusters had been selected, was divided in the same manner into 9 computing strata (numbered 35-43 in the table), each with 2 computing units.

In Domain [2], computing units were taken to be the same as the actual PSUs (localities). Strata were defined to include pairs of adjacent sample localities, following the systematic order of selection. This is the usual way of constructing 'collapsed strata' each with 2 units from a systematic sample of PSUs.

In this manner, 43 computing strata were defined as shown in Table 4B.(1), each with two computing replicates. The redefined design is very convenient for the application, especially of methods like the BRR, being based on the paired selection model. It also reflects the original sample structure in that the stratifying effect of systematic sampling is retained, apart of course from the usual overestimation of variance which the collapsed strata technique involves. By contrast, the random grouping of clusters and combining randomly across the implicit stratification provided by systematic selection does not bias the variance estimation, though it reduces its precision.

Sufficient information is not available to judge in quantitative terms the effect of such redefinition of the sample structure on the bias and precision of the variance estimates generated. Nevertheless, the illustration provides a useful example of the application of various techniques described in the preceding section.

TABLE 4B.(1).An example of the definition of computing units (effective primary selections) for the calculation of sampling errors.

(Source: Estudio Nacional de Salud, Bogota, Colombia).

stratum	clusters	computing unit	clusters	computing unit
01*	01,03,05,07,09,11,13	01	02,04,06,08,10,12,14	02
02*	15,17,19,21,23,25,27,29	03	16,18,20,22,24,26,28,30	04
03	(14)	05	(8)	06
04*	01,03,05,07,09,11,13 15,17,19,21,23,25	07	02,04,06,08,10,12,14 16,18,20,22,24,26	08
05	(20)	09	(19)	10
06	(18)	11	(16)	12
07	(18)	13	(17)	14
08	(12)	15	(14)	16
09	(21)	17	(18)	18
10	(15)	19	(18)	20
11	(16)	21	(24)	22
12	(18)	23	(16)	24
13	(16)	25	(18)	26
14*	01,03,05,07,09,11	27	02,04,06,08,10,12	28
15*	13,15,17,19,21,23	29	14,16,18,20,22,24	30
16*	25,27,29,31,33,35,37	31	26,28,30,32,34,36,38	32
17*	39,41,43,45,47,49,51	33	40,42,44,46,48,50,52	34
18	(10)	35	(12)	36
19	(6)	37	(10)	38
20	(18)	39	(16)	40
21	(20)	41	(14)	42
22	(16)	43	(16)	44
23	(14)	45	(18)	46
24	(17)	47	(14)	48
25	(16)	49	(16)	50
26*	01,03,03,07,09	51	02,04,06,08,10	52
27*	11,13,14,17,19	53	12,14,16,18,20	54
28*	21,23,25,27,29,31	55	22,24,26,28,30,32	56
29*	33,35,37,39,41,43	57	34,36,38,40,42,44	58
30	(8)	59	(8)	60
31	(26)	61	(6)	62
32	(19)	63	(14)	64
33	(17)	65	(14)	66
34	(18)	67	(13)	68
35*	01,03,05,07,09,11,13	69	02,04,06,08,10,12,14	70
36*	15,17,19,21,23,25,27	71	16,18,20,22,24,26,28	72
37*	29,31,33,35,37,39,41	73	30,32,34,36,38,40,42	74
38*	43,45,47,49,51,53,55	75	44,46,48,50,52,54,56	76
39*	57,59,61,63,65,67,69,71	77	58,60,62,64,66,68,70,72	78
40*	73,75,77,79,81,83,85,87	79	74,76,78,80,82,84,86,88	80
41*	89,91,93,95,97,99,101,103	81	90,92,94,96,98,100,102,104	82
42*	105,107.......115,117,119	83	106,108........116,118,120	84
43*	121,123.......131,133,135	85	122,124........132,134,136	86

NOTES. (*) indicates random groupings of clusters from the same `self-representing PSU' to form one or more pairs of computing units. In the remaining cases, the figure in parentheses indicates the number of clusters grouped to form a single computing unit.

4.6 SOFTWARE FOR VARIANCE ESTIMATION

The application of the variance estimation procedures described in Chapters 2 and 3 requires access to necessary computer facilities, especially to computer software of known quality and capability. The only procedure that may be applied as a part of the normal tabulation process is the independent replication method of Section 3.2.

Basic Requirements

The main advantage of general methods of variance estimation is that they can be applied to a wide variety of sample designs and types of statistics without modification to the basic procedure. This makes it possible to develop general purpose software for their application. Ideally, one should be able to perform large-scale, routine computations of sampling errors simply by specifying certain parameters for use by a suitable general purpose software for variance estimation.

Kaplan et al (1979) summarise some basic requirements which general programs of variance estimation from complex surveys should satisfy. They note that a general program "ideally should have great flexibility in dealing with various designs. The program should allow the user to describe his design exactly, accounting for strata, clusters, various stages of sampling, and various types of case weighting...If a program is to be of general use it must be reasonably convenient to learn and use. Such a program will not only be more useful, but will be easier to check and debug, and this, in turn, will improve accuracy. A good recoding system would allow for easy calculations of estimates for subpopulations. Missing value codes should exist and the program should be specific about its treatment of missing values, and small sample sizes (eg cluster sample sizes of zero or one)". Regarding the output generated, the authors note that it should "echo all the user commands: all options which were specified should be clearly repeated, including a description of the design. The labelling should be clear, and allow the user flexibility in naming his variables. The documentation of the output should be clear, concise and self-explanatory. It should also provide references which explain the statistical techniques programmed."

In large-scale applications a particularly important requirement for software is the ability to handle in an efficient manner the large number and variety of statistics for which sampling errors have to be computed. Efficiency refers not only to computing time, but also (and even more importantly) to the time and trouble required by the user in specifying the computations to be performed. In more specific terms, the following software features are desirable (Verma, 1982):

[1] The program should be able to handle, simply and cheaply, a large number of variables over different sample subclasses. It should not require the use of large computers or other very specialized facilities.

[2] In relation to the study of differentials between subpopulations, sampling errors for differences between pairs of subclasses should also be computed.

[3] It should be possible to repeat, in a simple way, the entire set of calculations for different geographical or administrative regions; such breakdowns are often required for substantive survey results.

[4] The computational procedure must take into account the actual sample design, in particular the effects of clustering and stratification, which influence the size of sampling errors. However, the program should not be limited to a particular sample design, such as a two stage design or the paired selection model.

[5] It should be able to handle weighted data.

[6] As far as possible, the program should not require any particular arrangement or form of input data. Where recoding of the raw input data is required, it is desirable that the software package itself should be able to handle this, without the need to write special programs for that purpose alone.

[7] In addition to calculating standard errors, it is also desirable that the program compute certain other derived statistics, such as coefficients of variation, design effects, and roh values. Such computed values may assist users to extrapolate to other variables and subclasses for the given sample and possibly also to future surveys. One of the objectives of calculating sampling errors is to provide information for sampling statisticians attempting to design other studies under similar survey conditions.

In-House Development Versus Acquisition of General Purpose Software

For organisations engaged in conducting diverse surveys on an ongoing basis, an important decision to be taken is whether to develop and maintain in-house its own special purpose software, or to try and acquire suitable general purpose software from some outside source. Several factors have to be taken into consideration in reaching a decision. The issue in the context of overall survey data processing has been considered in some detail in United Nations (1982; pp 99-114). The conclusion from the review is summarised as follows:

It is true that statistical offices in several developed countries have invested heavily in the development of general-purpose software systems for use on a wide variety of their statistical applications. These systems have been necessary to meet their specialized requirements for survey processing, as well as to integrate a standard data management philosophy across all applications. This is not to say that these organizations could not have been adequately served by already existing software packages but rather, in most cases, that the decision to develop reusable systems was based on a specialized need and the availability of high level programming staff to do the development.

Most developing country statistical offices do not have the luxury of having an abundance of high level programming staff to be able to contemplate development of specialized reusable software packages. These organizations are encouraged to use existing software systems available from other statistical offices or vendors, and to integrate them with smaller customized routines required for special needs, in order to avoid the need to program one-time-use customized systems. Writing customized software would seem like the more risky approach to take, as data processing staff turnover is usually quite high in developing countries, making the maintenance of customized software virtually unmanageable.

The above does not imply that there can be no problems in the choice and operation of appropriate software packages, or that such packages are available to meet all or most of the needs of a continuing survey programme.

The main advantage of in-house development is the potential ability of the software to meet specific requirements more effectively; while the main reason for seeking general purpose software from outside is usually the lower cost and time involved in its maintenance and use.

ILLUSTRATION 4C SOFTWARE FOR COMPUTING SAMPLING ERRORS: A REVIEW

This review of the available general purpose software for computing sampling errors in the context of complex, large scale surveys is presented in the form of an "illustration" for two reasons. Firstly, the review cannot be complete for lack of information: not all software is publicised, or is explicitly placed in the public domain. Secondly, the situation is prone to change and any information which can be provided here is of less long-term value than (hopefully) the other material included in this Technical Study.

A General Review of Available Software

The situation regarding software for variance estimation is rather different from that of software for general survey processing. For the latter, a variety of general-purpose programs are available, and often the user's main concern is to choose the one(s) most suited for his or her particular needs. However, in the area of variance estimation, the problem still is that very little software is available which is suitable for general use by different users. Of course there exist within individual organisations many special-purpose programs for variance estimation. However the vast majority of these are developed and maintained for the concerned organisation's own needs and specific applications, and are essentially non-portable and unsupported for use by other organisations or individuals.

For a particular user the suitability of any acquired general-purpose software depends on a variety of factors such as (i) particular requirements of the user, eg the type and diversity of sample designs encountered, the volume and sophistication of the computations required; (ii) hardware, software and 'personware' environment; (iii) flexibility, convenience, accuracy and reliability of the software; (iv) its portability to different settings; (v) how well the software is maintained by the supplying organisation; (vi) the degree of technical sophistication required for its use; and (vii) its cost, including the cost of initial purchase, maintenance and operation.

There are a small number of programs which could be considered suitable for general use. However a selection among even this small number is not easy. The available descriptions of software tend to be incomplete. It is difficult in particular to obtain reliable information on how portable and well supported any program really is. Claims made by suppliers in this regard are not always reliable, or at least may not be up-to-date. In any case the situation regarding software availability is constantly changing, and for this and other reasons, it is not possible in the present study to make definitive recommendations on particular programs. Wolter (1985) provides a description of 14 packages which were believed at the time of the review to be 'portable' and 'available' to some degree. Generally, the descriptions are not detailed or complete, and most cases appear to have been supplied by the developers themselves rather than being the result of an independent evaluation. The 14 programs mentioned above are listed in Table 4C.(1), with some brief remarks.

It can be seen from the table that even among this very limited number of programs, the great majority cannot be considered easily available or portable for general application. For example, some were not actually available at the time of the review (nos 1, 2, 9, 13 and 14 in the table below); some were limited to special applications and designs (4, 5, 6, 8); while some others were integrated with large systems not readily or cheaply available (6, 7, 10, 11). This leaves only PC CARP and CLUSTERS from the list. On the basis of this review the general conclusion regarding the availability of suitable variance estimation software to developing country survey organisations have to be rather pessimistic at this stage. Perhaps there are other suitable programs which have been missed in the review, or have been developed since. Nevertheless the above remarks underline the point that there is a dearth of good general purpose software for variance estimation which meet the basic requirements of flexibility, portability, reliability, ease in learning and using, good documentation, computing efficiency, low cost, and above all, active maintenance and support by the supplier.

TABLE 4C.(1). A review of general variance estimation programs.
(Source: Wolter, 1985)

Developer/Distributor	Remarks
01 **BELLHOUSE** D Bellhouse Univ of Western Ontario, Canada	Not yet available*.
02 **CAUSEY** Causey Bureau of the Census, USA	No longer available.
03 **CLUSTERS** V Verma International Statistical Institute	New portable PC version available since 1986; some support by ISI; also supplied by the DHS programme to participating countries.
04 **FINSYS-2** W E Frayer Colorado State University, USA	Specifically for forestry applications; restricted to certain designs.
05 **HESBRR** G K Jones National Centre for Health Statistics, USA	Developed specifically for Health Examination Survey; restricted to 2 PSUs per stratum designs.
06 **WASSTIM & WASSTVAR** D Morgenstein Westat Inc, USA	BRR restricted to 2 PSUs per stratum design; requires SAS.
07 **OSIRIS IV** L Kish Univ.of Michigan, USA	Expensive (also large annual renewal cost); not easily portable.
08 **PASS** D Thompson Social Security Administration, USA	Not portable; restricted to UNIVAC.
09 **RGSP** F Yates Rothamsted Experimental Station, UK	Versatile, but no longer formally supported.
10 **SPLITHAVES** J R Pryor Australian Bureau of Statistics	Part of ABS's Survey Facilities System; not portable.
11 **SUDAAN** V B Shah Research Triangle Institute,USA	Usable only in conjunction with SAS system.
12 **SUPER CARP** W Fuller Iowa State University, USA	New portable PC version available (PC CARP).
13 **U-SP** G B Wetherill University of Kent, UK	Not yet available*.
14 **VTAB & SMED83** National Central Bureau of Sweden	VTAB no longer distributed; SMED 83 still under development

*At the time of the review in 1985

In the following subsections, two of the more widely available and used programs (PC CARP and CLUSTERS) are briefly reviewed. Fuller details are available in users' manuals from the suppliers. Both programs are available for use on personal computers, free or at a nominal cost.

PC CARP: Cluster Analysis and Regression Program for Personal Computers

A description of the program is available in Fuller et al (1987).

PC CARP is available on IBM PC (AT or XT) and compatible computers with math co-processor and a minimum of 410K memory. The program is written almost entirely in FORTRAN (with a small portion in IBM Assembly language) and runs under DOS version 3.0 or higher. For variance estimation, the linearisation procedure (Ch. 2) is used. The program has a wide range of very useful analysis capabilities as summarised in the tables below reproduced from the above mentioned reference. However, before discussing these capabilities, it is important to note two limitations of the program in its present version. Firstly, in terms of sample design, the program has a limitation which may restrict its application in certain situations: presently it can be used to compute variances for one or two stage designs only. The second limitation is that the program assumes that there are no missing values, ie all data are available or have been imputed. Though a routine is provided for hot deck imputation of missing values, this restriction is inconvenient. Missing values occur in all surveys and in many situations it is more appropriate (and simpler) to exclude items with missing values from the computations, than to always have to impute them using a more or less arbitrary procedure.

Turning to Table 4C.(2) summarising analysis capabilities of the program, 'Population Analysis' refers to the computation of sampling errors for aggregates, ratios and differences of ratios for the total sample. This includes standard errors, design effects and, except for differences of ratios, relative standard error and covariance matrix of any specified set of estimates. 'Stratum Analysis' refers to the decomposition of variance into within and between stratum components. 'Subpopulation Analysis' means computation of sampling errors for subclasses defined by crossing two or more classification characteristics. Sampling errors may be obtained for any set of estimates under the classification structure so defined; however the full covariance matrix is not produced.

The above capabilities meet the common requirements of sampling error computations in large-scale descriptive surveys for complex statistics up to ratios, and in some cases, differences of ratios. A more distinguishing feature of the program is the provision for some other analyses which include the following.

[1] Two Way Table Analysis refers to analysis by two classification variables and a dependent variable. Tables of cell totals, of proportions based on row totals, and of proportions based on the grand total are computed for each dependent variable specified. Standard errors are computed for all estimators and a test statistic for the hypothesis of proportionality is obtained.

[2] Regression Analysis refers to the computation of weighted least squares regression coefficients, and an estimated variance-covariance matrix which takes into account the sample design. Multiple degrees of freedom F-tests for sets of coefficients and the usual test statistics are available, as is the option of obtaining residuals and predicted values.

TABLE 4C.(2). Main features of PC-CARP.
(Source: the distributor)

Versions of PC CARP

Machine	**Form of PC CARP**	**Required Memory**
IBM PC/AT with Math Co-Processor	1 High density diskette	450K
IBM Personal Computer with Math Co-Processor	2 Double density diskettes	410K
IBM PC/XT with Math Go-Processor	2 Double density diskettes	410K

Analysis capabilities of PC CARP

Analysis	**Coeff. of var**	**Cov. matrix**	**Design effect**	Comments
Population Analyses				
Total Estimation	x	x	x	50 variables maximum
Ratio Estimation	x	x	x	50 variables maximum without covariances, 15 with covariances
Difference of Ratios	x		x	15 variables maximum
Stratum Analyses				
Totals	x		x	50 variables maximum
Means	x		x	50 variables maximum
Proportions	x		x	50 variables maximum
Subpopulation Analyses				
Totals	x		x	Crossed classif. Multiple variables
Means	x		x	Crossed classif. Multiple variables
Proportions	x		x	Crossed classif. Multiple variables
Ratios	x		x	Crossed classif. Multiple variables
other Analyses				
Two-Way Table		x		50 cells maximum, proportionality test
Regression		x		50 variables maximum Multiple d.f. tests Y-hat, residuals
Univariate			x	Multiple variables, empirical CDF, quantiles

[3] Univariate Analysis provides statistics describing the distribution of a variable over a specified subpopulation. Estimates of the mean, variance, distribution function, quantiles and interquantile range are produced.

[4] Some additional features include the following: incorporation of the finite population correction (assuming uniform or averaged sampling rates within strata); estimation of quantiles and their standard errors; estimation of a multivariate logistic model using an iterative least squares algorithm; and handling of post-stratification where the estimates have been adjusted to match known population control totals.

CLUSTERS: A Package Program for the Computation of Sampling Errors for Clustered Samples. Version 3

A description of the program is available in Verma and Pearce (1986).

In terms of computer requirements, CLUSTERS is similar to PC CARP. It is available on IBM PC and compatibles with math co-processor, is FORTRAN based, and uses the linearised method for variance estimation. The program was originally developed in the mid-Seventies and over the years has been made available freely on request to statistical organisations and individual researchers in many developing and developed countries. The new Version 3 was developed in 1985-86 to make the program more flexible and easier to use in certain respects, and above all, to make it available on personal computers. Among other uses, the program was used on a large scale to compute sampling errors for all surveys in developing countries conducted under the World Fertility Survey programme (1972-1984); and more recently it has been used systematically in developing country surveys conducted under the Demographic and Health Surveys programme. The package has also been used by countries for computing sampling errors for labour force and other household surveys.

The program is more limited than PC CARP in the range of analyses performed. Instead, it is focused on the primary task of large-scale computation of sampling errors for diverse statistics over various subclasses and domains of the sample. It includes flexible facilities for specification of parameters of the sample design relevant to the computations, and also provides a useful set of recoding facilities for defining statistics and subclasses for which sampling errors are to be computed. It handles missing values by appropriately excluding them from the computations. Several features of the program which make it particularly suited for large scale computation of sampling errors in descriptive surveys with complex designs are noted in Table 4C.(3).

TABLE 4C.(3). Some useful features of clusters.
(Source: Verma, 1982)

1. Sample structure. The program computes sampling errors taking into account the actual sample design, in particular clustering, stratification and weighting of the sample. There are no specific restrictions on the design, except for the basic ones for the linearisation method, namely at least two independent primary selections per stratum with replacement. (A limitation of the program is that the finite population correction is always disregarded.) The program includes flexible facilities for specifying computing strata and primary units, as well as specifying sample weights.

2. Data input and transforation. The program is designed to minimise the need to restructure or modify the data prior to their use in the program. It has been extended to handle hierarchical data files which may contain records for units at different levels such as households and individual persons. Data files may optionally be described by an associated file called the 'dictionary'; the program accesses the dictionary for information on the variables. It is often necessary to recode input data before the required statistics can be computed. For this purpose the program includes a set of useful recoding facilities which can define new variables by combining or transforming one or more existing variables, exclude cases not belonging to specified categories (subclasses), identify and deal with missing values, etc.

3. Handling diverse variables and subclasses. The program allows the specification of a set of variables and a set of subclasses and then automatically proceeds to compute estimates and sampling errors for the whole 'variable by subclass' matrix, ie for each variable over the full set of subclasses specified. This is convenient because in survey analysis often the same system of classification is relevant to all (or most) survey variables. This feature reduces the work required in specification of the computations to be performed. In computing sampling errors for a subclass, the program also computes sampling error for the dichotomous variable defining the subclass, treating it as a characteristic distributed over the sample.

4. Subclass differences. The sample subclasses for which sampling errors are to be computed can be specified in pairs. In that case CLUSTERS automatically calculates the difference and its standard error for each subclass pair. A given subclass may, if desired, appear in more than one pair; moreover the subclasses in a pair need not necessarily be non-overlapping or exhaustive.

5. Separate results for geographical domains. The entire set of calculations for variables over sample subclasses and for differences between subclass pairs can be repeated for separate geographical domains in to which the survey universe may have been divided. This repetition is extremely straightforward from the user's point of view and does not involve much additional computer time. One restriction regarding this facility in CLUSTERS is that the geographical regions must be non-overlapping and the sample must be selected independently within each region.

6. Derived statistics. In addition to standard errors, the program produces related statistics such as relative error, 95% confidence intervals, standard deviation, coefficient of variation, design effect (deft) and rate of homogeneity (roh).

7. Type of estimators. The program is confined to the computation of errors for 'descriptive' statistics including proportions, percentages means and ratios of pairs of substantive variables. Differences of ratios of the same characteristics defined over different subclasses are handled. These cover most types of statistics commonly encountered in large-scale household surveys.

The main limitations of the program are that (i) it is not designed to provide directly information on variance components; (ii) it does not handle more complex statistics than ratios and differences of ratios, statistics such as double ratios or regressions which may be of interest in some surveys; (iii) it has no provision for more sophisticated analyses, other than the basic task of efficiently computing sampling errors for descriptive statistics on a large scale in multisubject surveys; and (iv) an unnecessary but inconvenient limitation of the program in the present form is that it does not output sampling errors of simple aggregates.

PART II
ANALYSIS

5

DECOMPOSITION OF THE TOTAL VARIANCE

5.1 INTRODUCTION

The concern in the discussion of the variance estimation procedures in Part I has been with estimating the total sampling variance. This concern is correct because in practical survey work the first priority must be given to computing overall variance of the diverse estimates produced. This information is essential for proper interpretation and use of sample survey results. Analysis of the total variance into components is a more complex and demanding task. Nevertheless, it is necessary for survey design work to isolate, to the extent possible, components of the overall variance which can be related to important features of the design and to various stages of selection and estimation.

Detailed consideration of the complex topic of variance components is beyond the scope of this Technical Study. This chapter considers the following selected aspects, which are important in the analysis and use of the information on sampling errors:

[1] Decomposition of the total variance of a survey estimator into the overall effect of the design (as measured by the design effect, $deft^2$), and what the variance would have been with a simple random sample of elements of the same size.

[2] Decomposition of the design effect into the contribution of haphazard or random weights, and the overall effect of other complexities in the design.

[3] Identification of the effect of certain steps in the estimation procedure (post-stratification, ratio adjustments, composite estimation etc) on the variance of the resulting statistics; in particular, of the effect of variability of external weights, when the actual weighting factors applied depend on the results obtained in the particular sample.

[4] Assessment of at least the approximate contribution of various sampling stages and the effect of stratification at various levels.

[5] And a more formal analysis of the components of total variance by sampling stage in multistage designs.

In the above, [1] and [2] are no doubt the most important tasks as concerns practical survey work. Fortunately they are also the ones most easily accomplished. Simple procedures are available for estimating the equivalent SRS variance, from which deft can be computed given the variance of the actual sample. Simple and robust procedures are also available for separating out in the design effect the contribution of random (but fixed) weights. Indeed, design effects are routinely produced by most general purpose variance estimation programs such as CLUSTERS and PC CARP described in Chapter 4.

Objective [4] can also be accomplished by using procedures similar to those for computing the overall design effect, [1] - though here the procedures tend to be somewhat more approximate, more specific to details of the sample structure, and more demanding in terms of the computational work involved. The linearisation method of variance estimation described in Chapter 2 is perhaps the most suited for this type of analysis; this may also be the case for objective [5].

Objective [3] can be important when the estimation procedures involved are complex and demanding in terms of the time and effort required for their application. Generally, the repeated replication procedures described in Chapter 3 are better suited for this type of analysis.

Decomposition of the total variance by sampling stages, [5], is usually more difficult and complex. There are several problems in estimating variance components by sampling stage. The decomposition of overall variance into components involves complex procedures and greatly increased computational work, and often the results obtained are numerically unstable. Also, by their very nature, the computational procedures have to be more specific to detailed features of the sample design, making it difficult to establish common procedures applicable to different designs encountered in practice.

5.2 THE DESIGN EFFECT, AND VARIANCE IN AN EQUIVALENT SRS

The standard error, $se(\bar{y})$, for a statistic (say a mean, $\bar{y}$) estimated from a complex sample is factored into two parts each of which is discussed below:

$sr(\bar{y})$, the standard error which would have been obtained in a simple random sample of the same size;

deft, the design effect, defined as the ratio of the standard error for the actual design, to that for a simple random sample of the same size:

$$deft = \frac{se(\bar{y})}{sr(\bar{y})} \tag{5.1}$$

5.2.1 THE DESIGN EFFECT

The design effect (deft) is a summary measure of the effect of departures of the actual sample design from simple random sampling of elements. It is a comprehensive measure which attempts to summarise the effect of various complexities in the design, especially those due to clustering, stratification and weighting. It may also incorporate the effects of ratio or regression estimation, double sampling, variable sampling rates, etc. In practical survey work, departures from simple random sampling are introduced to reduce the cost of and improve control over field operations. These benefits have to be weighed against the loss in sampling efficiency measured by deft. Deft is one of the most commonly used and useful measuresof efficiency of the sampling design; many samplers include it as a routine item in the output of variance computation. Examples of defts from a number of developing countries were given in Illustration 4A.

Since deft itself is a measure incorporating the effect of various features of the design and estimation procedure, it may be decomposed further into components reflecting specific aspects of the design, such as the effect of weighting, estimation procedures, cluster sizes and other features of the sample structure. These issues will be considered in the following sections and in Chapter 6.

5.2.2 VARIANCE IN A SRS OF THE SAME SIZE; POPULATION VARIANCE

To compute the design effect, it is necessary to estimate

[1] the variance under the actual design. Various practical procedures for this have been described in Part I.

[2] the variance, for a given survey estimator, which would pertain in a simple random sample of the same size.

How can [2] be accomplished, when the actual observations we have available are from the actual complex design, rather than from a simple random sample of elements? Here is a point of great practical relevance: In most practical samples the equivalent SRS variance can be estimated well and simply from the sample observations. This is based on a remarkable result of sampling theory that, from the results of a given complex sample, we can estimate what the sampling error would have been in the hypothetical situation if certain complexities had not been present in the actual design. The procedure is simply to apply the computational method by assuming that the complexities concerned were not present. The sampling error for an equivalent simple random sample of the same size - and hence the overall design effect - can be estimated from the complex sample by applying to it the ordinary SRS variance estimation formulae (see Technical Note at the end of this section for some further discussion of the procedures). While this is the clearest and most common application of the procedure, the idea can be extended to explore the effect of particular features of the design; some possibilities are described in Section 5.5 and Illustrations 5C and D.

Some expressions for SRS Variance

We will first consider the estimation of equivalent SRS variance for a complex, but self-weighting ('epsem') sample.

[1] Means and Proportions.

Assuming a simple random sample of size n with replacement, we have the well-known expression for the variance of a sample mean:

$$var(\bar{y})_o = sr^2(\bar{y}) = \frac{s^2}{n} \tag{5.2}$$

where the subscript 'o' is used to indicate simple random sampling, and

$$s^2 = \frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n-1} \tag{5.3}$$

estimates the variance between individual elements in the population

$$S^2 = \frac{\sum_{j=1}^{N} (Y_j - \bar{Y})^2}{N-1} \tag{5.4}$$

S^2 is called the population variance. Its square root, S (or its estimate s), is the standard deviation, as distinguished from standard error, which measures the variability of the sample estimator rather than of individual elements in the population. S^2 does not depend on the structure of the sample, but only on characteristics of elements in the population. As mentioned above, in most practical samples, it can be estimated well and simply from the sample observations irrespective of the complexity of the design, except for the effect of sample weights, as noted below.

(For simplicity, it is assumed in this section that the finite population correction is inapplicable, negligible, or can be appropriately introduced in the expressions given here.)

For a proportion p, the expression for s^2 takes the well-known form

$$\begin{aligned} s^2 &= \frac{n}{n-1} p.(1-p) \approx pq; \\ var(p)_o &= \frac{s^2}{n} = \frac{p.(1-p)}{n-1} \approx \frac{pq}{n}; \quad with\ q = 1-p. \end{aligned} \tag{5.5}$$

[2] Ratios

The concept of population variance can be extended to other, more complex types of estimators as well, in so far as it can be expressed in a form such as (5.4) involving quantities (like y_j) defined at the level of individual elements (j). For example, for a ratio R = Y/X, we have

$$S^2 = \sum Z_j^2/(N-1)$$

where Z_j is an auxiliary variable defined at the level of individual units as

$$Z_j = (Y_j - R.X_j)/\bar{X}$$

With r as the estimate of R from a sample of size n, S^2 is estimated from the sample values as

$$s^2 = \sum z_j^2/(n-1), \quad with\ z_j = (y_j - r.x_j)/\bar{x} \tag{5.6}$$

giving the SRS variance with ratio estimate as $var(r)_o = s^2/n$, as before. In the above,

$$\bar{X} = X/N \ ; \qquad \bar{x} = x/n$$

are average values of the denominator in the ratio, for the total population and the sample respectively.

[3] Differences between Samples or Subclasses

Extension to differences between non-overlapping populations sampled independently is straightforward, but somewhat more complex in the presence of overlaps. This is discussed below in terms of differences of proportions.

If the proportions (p and p') between two mutually exclusive groups (of size n and n' respectively) are being compared, variance of the difference is simply the sum of their individual variances

$$var(p-p')_o = \frac{p.(1-p)}{n-1} + \frac{p'.(1-p')}{n'-1}$$

An example of the above is the comparison of proportions of poor in two (mutually exclusive) socio-economic groups. Sometimes the interest is in comparing proportions according to two characteristics in the same base population. If the two proportions (numerators) are mutually exclusive - for instance the proportion of voters voting for party A, and the proportion voting for party A' - then:

$$var(p-p')_o = \frac{[(p+p') - (p-p')^2]}{n-1}$$

where n is the size of SRS drawn from the common base population to estimate the proportions. This follows from the observation that

$$var(p-p')_o = var(p) + var(p') - 2.cov(p,p')$$

and for mutually exclusive proportions with a common base, cov(p,p') = -p.p'/(n-1).

When the two proportions overlap, the above is modified to

$$var(p-p')_o = \frac{[(p+p'-2.p'') - (p-p')^2]}{n-1}$$

where p" is the overlap between proportions p and p'.

Weighted Samples

When the actual sample observations have been weighted, the expressions for estimating the variance (without the effect of weights on it) for a simple random sample of the same size have to be modified with the weights. With w_j as the weights for individual elements (j), we have for example for a mean

$$\bar{y} = \frac{\sum_j w_j.y_j}{\sum_j w_j}; \quad s^2 = \frac{n}{n-1}.\frac{\sum_j w_j.(y_j-\bar{y})^2}{\sum_j w_j} \tag{5.7}$$

For a proportion p, the above expression applies, with y_j defined appropriately as a dichotomous variable p_j (=0 or 1):

$$p = \frac{\sum_j w_j.p_j}{\sum_j w_j}; \quad s^2 = \frac{n}{n-1}.\frac{\sum_j w_j.(p_j-p)^2}{\sum_j w_j} = p.(1-p)$$

Note that the form of s^2 for a proportion is identical to that for the unweighted case; the only difference is that p itself is estimated with the weights.

More generally, for a ratio we have

$$r = \frac{\sum_j w_j.y_j}{\sum_j w_j.x_j}; \quad s^2 = \frac{n}{n-1}.\frac{\sum_j w_j.z_j^2}{\sum_j w_j} \tag{5.8}$$

with

$$z_j = \frac{y_j - r x_j}{\bar{x}}; \quad \bar{x} = \frac{\sum_j w_j x_j}{\sum_j w_j}$$

Expressions of the above form for s^2 can also be applied to other more complex statistics with z_j appropriately defined. This extension is based on the linearisation procedure described in Chapter 2.

A most important point is that though weights appear in the above expressions for s^2, the quantity it estimates is the (unweighted) population value S^2, and hence the sampling variance estimated by s^2/n still refers to a selfweighting simple random sample of the same size (n) as the actual sample. The effect of weighting is of course incorporated into the variance for the actual design computed using the procedures described in Chapters 2 and 3, and hence into the design effect defined from the above. Separation of the effect of weighting from the overall design effect is discussed in Section 5.3.

A Note on Terminology

By convention, the term 'design effect' is used both for the ratio of actual to SRS variances, and for its square-root, ie the ratio of the standard errors. To avoid confusion where necessary, we use 'deft^2' when the reference is to the ratio of variances, and 'deft' for the ratio of standard errors. 'Deff' has also been used in place of deft^2, though now this usage is less common; however, a subtle distinction has sometimes been drawn between the two in the sense that the denominator in the case of deff does not include the finite population correction (fpc), while that in deff^2 does. The numerator (the variance of the actual sample) in either case is meant to include the fpc appropriately.

Some authors have preferred to reserve the term 'design effect' for the ratio of variances, and have used the term 'design factor' to refer to the ratio of standard errors.

5.2.3 TECHNICAL NOTE ON ESTIMATING SRS VARIANCE FROM A COMPLEX DESIGN

The objective of this note is to give a clearer understanding of the procedure described above for estimating the SRS variance from data obtained from a complex sample. Its basis follows from what has been called the 'argument of symmetry' (Cochran, 1973; Sec. 2.3). Consider for example a self-weighting sample of any design. Since each element in the population has the same chance of being selected into any sample, every unit in the population appears exactly the same number of times when the collectivity of all possible samples is considered. This implies that the mean per element of all possible samples (which by definition is the expected value of the sample estimator) is the same as the mean per element in the population, giving the well-known result that the sample mean provides an unbiased estimate of the population mean:

$$E(\bar{y}) = \bar{Y}$$

The point is that the above applies not only to a particular variable, but to any variables defined in terms of individual values, such as $z_j = y_j, y_j^2, y_j.x_j$, etc, but without involving cross-products of values for different elements. (The argument of symmetry does not apply to cross-products, because complexity of the design affects the probability with which any particular combination of elements appears in the samples.) The above also applies to quantities like

$(y_j - \bar{Y})^2$ or $(y_j - Rx_j)^2$, where $\bar{Y}$ and R are constant (population) parameters. With a reasonably large sample size, it also applies with generally only a slight approximation to quantities like $(y_j - \bar{y})^2$ or $(y_j - rx_j)^2$, with the population parameters replaced by their sample values. On this basis s^2 estimates S^2, ie $E(s^2) = S^2$, irrespective of the complexity of the design. It can be established that the actual relationship is

$$E(s^2) = \frac{n}{n-1} \cdot \left[\frac{N-1}{N} . S^2 - Var(\bar{y})\right] \tag{5.9}$$

in which the second term on the right is of the order of (1/n) compared to the first, ie usually negligible by comparison. For a simple random sample without replacement, we have the equality $E(s^2)=S^2$ since $Var(\bar{y})=(1-\frac{n}{N}).\frac{S^2}{n}$. In a complex design, s^2 slightly underestimates S^2 in so far as the complexity increases the negative variance term on the right hand side.

The argument is easily extended to the general case of a complex sample when the individual elements are weighted inversely proportional to their respective probabilities of selection. The argument of symmetry goes as follows. If p_j is the probability of selection of an element j, then the number of times it appears when all possible samples are considered is proportional to p_j. Now if the contribution of each element to all possible samples (ie, to the expected value) is divided by its p_j (ie, multiplied by its inverse, w_j), the result is that the "effective" number of appearances when all possible samples are considered is the same for all elements in the population. This means that any functions of individual values of the type noted above, which involve individual values but not cross-products across different elements, is weighted by w_j to estimate the corresponding population parameter. From this expressions like (5.7) and (5.8) follow.

5.3 THE EFFECT OF UNEQUAL WEIGHTS

In many surveys the objective is to produce estimates at various levels of aggregation such as at the national as well as subnational levels. Comparisons among subnational estimates are also required. The different objectives results in conflicts requiring compromises in sample allocation. For any given objective, the compromise allocation essentially represents "random" weighting, the effect of which is to inflate the variance for that objective. The important thing is that unequal weights tend to affect (inflate) the variance of all estimates for different variables over different subclasses in a rather uniform way, independently of the structure of the sample except for weighting itself. Herein lies the practical utility of isolating this effect. Its magnitude has been expressed in very simple equivalent forms (Kish 1965, p427, and 1989, p183; the original author uses the symbol (1+L) for the quantity denoted by D_w^2 here):

$$D_W^2 = \sum_h (W_h.w_h).\sum_h (\frac{W_h}{w_h}) = \frac{n.\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2} = n.\sum_h \frac{W_h^2}{n_h} \tag{5.10}$$

$$with \quad n = \sum_h n_h; \quad \sum_h W_h = 1$$

D_w is the factor by which standard error is inflated due to random weighting. In the above, w_h are the weights, uniform for the n_h units in stratum h; and W_h are the relative sizes of the strata in the population. The three forms in the equation are equivalent since weights w_h are generally taken to be inversely proportional to the sampling rates, $f_h = n_h/N_h = n_h/(N.W_h)$.

The above can also be written in terms of the coefficient of variation of the weights as

$$D_W^2 = \frac{n.\sum_j w_j^2}{(\sum_j w_j)^2} = 1 + cv^2(w_j) \tag{5.11}$$

where

$$cv^2(w_j) = \frac{1}{n.\bar{w}^2}.\sum_j (w_j - \bar{w})^2; \quad \bar{w} = \frac{\sum_j w_j}{n}$$

A more precise expression for the loss factor estimated for a ratio r=y/x and with weights varying generally at the level of individual units is

$$D_W^2 = \frac{n.\sum_j (w_j z_j)^2}{(\sum_j w_j z_j)^2} \tag{5.12}$$

where

$$r = \frac{y}{x} = \frac{\sum_j w_j.y_j}{\sum_j w_j x_j}; \quad z_j = \frac{1}{x}.(y_j - r x_j)$$

Illustration 5A demonstrates the computation of D_W, the design effect due to weighting using (5.10). Illustration 6F gives several examples showing that this effect tends to be similar for diverse variables and subclasses.

ILLUSTRATION 5A THE EFFECT OF ARBITRARY WEIGHTS ON VARIANCE

Table 5A.(1) shows the relative weights applied in samples from some developing countries. (The example is taken from the Demographic and Health Surveys programme.) Each sample was essentially self-weighting within each of a number of major geographical domains in the country, but weighted across the domains. The table shows the sample size (n_h) and the relative weight (w_h) for each domain. (For convenience, the weights have been scaled such that the average value is 1.0.) The last two columns show the increase in the sampling error due to weighting for estimates produced at the national level, computed on the basis of equation (5.10). These factors apply essentially unchanged to all variables estimated at the national level, as well as to estimates over cross-classes which are distributed across the geographical domains. For estimates produced at the domain level, there is of course no effect due to weighting because the samples are self-weighting within domains.

TABLE 5A.(1) Computing the effect of arbitrary weights on variance.
(Source: Aliaga and Verma, 1991)

	weighting domain								Loss factor	
	1	2	3	4	5	6	7	8	D_w^2	D_w
UGANDA										
n_h	964	128	689	1108	132	1289	420	-	1.16	1.08
w_h	0.56	1.83	1.68	0.94	1.83	0.97	0.62	-		
BOTSWANA										
n_h	2258	2110							1.19	1.09
w_h	0.58	1.45								
DOMINICAN REPUBLIC										
n_h	1336	631	1302	891	926	758	1016	789	1.35	1.16
w_h	2.09	0.71	1.39	0.91	0.43	0.66	0.55	0.45		
BRAZIL										
n_h	749	769	847	1029	1794	709			1.12	1.06
w_h	0.82	1.69	1.35	0.89	0.88	0.48				
EGYPT										
n_h	272	246	116	279	286	8621			1.05	1.03
w_h	0.27	0.22	0.35	0.56	0.55	1.08				

Table 5A.(2) shows the effect of weights on the overall (total sample) defts. The first panel shows the actual defts, including the effect of weighting. The values have been averaged over groups of substantively similar variables. The variables and the groups were defined in Table 4A.(3) in the previous chapter. Deft values for individual variables was shown in Table 4A.(4). As explained in Chapter 6, such averaging is often necessary and useful.

The second panel of the table shows what the total sample design effects would have been in the absence of weighting which inflates the variance; that is, the figures shown are the actual deft values from the first panel, divided by the loss factor D_w for the design from Table 5A.(1). For certain countries the values in the two panels are the same because the samples involved were self-weighting; these are the countries shown in this table, but not in 5A.(1).

TABLE 5A.(2). Examples of the effect of weighting on deft values.

(i) Deft values averaged over groups of variables

country	GHANA	ZIMBABWE	UGANDA	BOTSWANA	DOMINICAN REP	PERU	BRAZIL	EGYPT	ECUADOR	SENEGAL	KENYA	THAILAND
var	1	2	3	4	5	6	7	8	9	10	11	12
group												
1	1.15	1.04	1.19	1.25	1.28	1.12	1.26	1.40	1.26	1.32	1.46	1.72
2	1.33	1.19	1.23	1.51	1.31	1.12		1.51	1.64	1.23	1.35	1.65
3	1.57	1.22	1.14	1.34	1.21	1.20	1.31	1.42	1.08	1.34	1.59	1.71
4	1.84	1.20	1.52	1.87	1.27	1.34	1.24	1.91	1.65	1.61	2.00	1.82
5	1.31	1.09	1.17	1.25	1.10	1.07	1.15	1.23	1.14	1.16	1.30	1.37
6	1.56	1.58	1.69	1.64	1.86	1.40	2.28	2.76	1.69	2.02	2.01	2.10
all	**1.41**	**1.18**	**1.29**	**1.48**	**1.30**	**1.18**	**1.29**	**1.67**	**1.37**	**1.39**	**1.55**	**1.69**
D_w	**1.00**	**1.00**	**1.08**	**1.09**	**1.16**	**1.00**	**1.06**	**1.03**	**1.00**	**1.00**	**1.28**	**1.22**

(ii) Deft values after removing the effect of arbitrary weights

group	1	2	3	4	5	6	7	8	9	10	11	12
1	1.15	1.04	1.10	1.15	1.10	1.12	1.19	1.36	1.26	1.32	1.15	1.41
2	1.33	1.19	1.15	1.39	1.13	1.12		1.47	1.64	1.23	1.06	1.36
3	1.57	1.22	1.06	1.23	1.05	1.20	1.24	1.38	1.08	1.34	1.24	1.40
4	1.84	1.20	1.41	1.71	1.10	1.34	1.17	1.87	1.65	1.61	1.57	1.49
5	1.31	1.09	1.08	1.15	0.94	1.07	1.08	1.20	1.14	1.16	1.02	1.13
6	1.56	1.58	1.57	1.50	1.60	1.40	2.16	2.69	1.69	2.02	1.58	1.73
all	**1.41**	**1.18**	**1.20**	**1.36**	**1.12**	**1.18**	**1.21**	**1.63**	**1.37**	**1.39**	**1.22**	**1.39**

5.4 THE EFFECT OF VARIABILITY IN THE ESTIMATION WEIGHTS

The issue of variable weights should be distinguished from essentially random but fixed weights considered in the preceding section.

Sample data may be weighted for various reasons in estimating the population parameters. Apart from differences in selection probabilities, weighting may be introduced for nonresponse, post-stratification or ratio adjustments. While usually the main objective of weighting is to control bias, weights which randomly depend on the particular units which happen to be selected may also affect the variance of the estimates. For instance, post-stratification or ratio weights often reduce the variance while nonresponse adjustment weights tend to inflate it. The effect of weights treated as constants is automatically incorporated into the variance estimation procedures described in Part I. However the treatment of weights variability in variance estimation requires special considerations.

The method of linearisation is difficult to adapt to take into account the contribution of variable weights. The repeated replication methods are more suited for the purpose. In the production of replicated estimates, the weights may be introduced in one of the following two forms:

[1] a common set of weights computed from the full sample and applied to all the replications; or

[2] weights computed and applied separately for each replication, using a common procedure.

Procedure [2] takes into account the effect of weight variability, but [1] does not. Their comparison will show the importance of this effect. Ideally [2] should be used in variance estimation, but it can be much more costly in terms of the computations involved. When the effect of weight variability is not important, the more economical procedure [1] will suffice. In that case, of course, the linearisation method can also be used in place of repeated replication, unless the latter is preferred in view of the complexity of the statistics involved. In practice it is also possible to use a combination of [1] and [2]: using procedure [2] for steps in the estimation procedure where the contribution of weight variability is important and/or can be handled without too much additional computational work, and using [1] in the remaining steps. Several examples of the approach are given in the following illustration.

ILLUSTRATION 5B MAGNITUDE OF THE EFFECT OF VARIABLE WEIGHTS

Several investigations have shown that in many situations the effect of weight variability on variance is not important. For example Kish and Frankel (1970) computed variances of several types of statistics using the BRR method. Some results are shown in Table 5B.(1). Column (4) in the table shows the factor by which the standard error is changed when the effect of weight variability is taken into account. The factors are mostly close to 1.0, indicating a mere 2-3% increase in standard error or 5-6% increase in variance due to weight variability. Similarly small effects have been reported by Rust (1987), as shown in Table 5B.(2). However, the same author also reports a case with striking reduction in the estimated variance when weight variability was taken into account by computing the weights separately for each replication, using in this example the BRR method. Some results along with a brief commentary are shown in Table 5B.(3). The figures for the linearisation method in the table do not take this effect into account. The differences between the two estimates are marked. Such strong effects of weight variability may be present when the post-stratification variables are highly correlated with the substantive variables being estimated, so that the use of 'correct' weights in each replication makes a significant difference to the precision of the estimates produced.

TABLE 5B.(1). An example of negligibly small effect of the variability in weights.
(Source: Kish and Frankel, 1970)

AVERAGE VALUES OF $\sqrt{DEFF}$ FOR SEVERAL STATISTICS FROM 16 REGRESSIONS, WITH DIFFERENT PREDICTANDS FROM SAME 3 PREDICTORS

Statistic	(1)[a]	(2)[b]	(3)[c]	(4) = (2) ÷ (3)
Ratio means	18	1.7998	1.7549	1.0256
Simple correlations	51	1.2616	1.2802	0.9855
Partial correlations	48	1.3995	1.3487	1.0377
Multiple R	16	1.4653	1.4217	1.0307
Regression coefficients	48	1.2948	1.2668	1.0221

[a] (1) =number of different statistics averaged in Columns (2) and (3).
[b] (2) =average values of $\sqrt{\text{deff}}$ with correct weighting system.
[c] (3) =average values of $\sqrt{\text{deff}}$ with approximate weighting system.

TABLE 5B.(2). Another example of small effect on variances of the variability in sample weights. (Source: Rust, 1987)

BRR estimates of coefficients of variation for the Title IV Quality Control Study - Pell Grant Awards

Parameter	Subgroup	Incorporating sampling variation in weights	Ignoring sampling variation in weights
Total absolute error	All students	.107	.117
	Indep. students	.289	.280
	Dep. students	.082	.094
	Pell Grant only	.235	.237
Total overpayment	All students	.060	.073
	Indep. students	.116	.122
	Dep. students	.059	.070
	Pell Grant only	.137	.146
Mean error per student with error	All students	.110	.110
	Indep students	.290	.277
	Dep. students	.080	.080
	Pell Grant only	.214	.210
Proportion with error	All students	.030	.029
	Indep. students	.046	.047
	Dep. students	.034	.033
	Pell Grant only	.056	.055

TABLE 5B.(3). An example of large effect on variances of the variability in sample weights.
(Source: Rust, 1987)

Comparison of BRR and "simple" linearization estimates of design effect for Hispanic HANES (from Lago *et al.* (1987)).

Subgroup	Statistic					
	Mean weight		Mean height		Mean cholesterol	
	BRR	Linearization	BRR	Linearization	BRR	Linearization
All persons	0.64	2.31	1.20	3.59	0.65	1.49
All males	0.47	1.38	0.58	2.20	0.96	1.14
All females	0.56	1.93	1.01	3.03	0.63	1.03
Male 45-54	0.60	0.59	1.66	1.36	1.68	1.66
Female 25-34	0.63	0.60	2.78	2.78	1.16	1.16

As part of a broader study, a similar investigation has been undertaken for estimates for the Mexican American component of the Hispanic Health and Nutrition Examination Survey (HHANES) conducted for the U.S. National Center for Health Statistics (Lago *et al.* (1987)). Variance estimates using BRR, incorporating sampling variation in poststratification weights, were compared with those obtained from linearization, with the effect of poststratification on variance ignored. Poststrata were formed on the basis of age and sex. Little difference was found for most parameters, but for three statistics, mean weight, mean height and mean cholesterol level, the estimated design effects from linearization were several fold those from BRR for whole population estimates (see Table). These variables (weight, height and level of cholesterol) are highly correlated with age and sex, the poststratification variables, so that the use of poststratification gave rise to considerable reduction in sampling variance. In failing to reflect this, the method of linearization used gave gross overestimates of sampling error. The table shows that the differences between the two methods disappear for estimates for specific age-sex cells, where the use of poststratification has no effect on the precision of estimation.

5.5 EXPLORING THE EFFECT OF SAMPLING STAGES AND STRATIFICATION

While accurate computation of variance components by stages in a multistage design can be complex (as considered in the next section), the general methods described in Section 5.2 can be adapted to yield reasonable approximations in many situations for the effect of certain features of the design, such as that of one or more of the highest stages of sampling or that of stratification at various levels.

This approach is based on the result of sampling theory (noted in Sec 5.2.3) that, from the observations from a given complex sample, we can estimate what the sampling error would have been in the hypothetical situation with some complexities of the design removed. For example to investigate the effect of the first stage in a multi-stage design, the procedure is to apply the computational method by assuming that the highest stage concerned was not present, that is, as if the second stage units in the sample had been selected directly as the PSUs. Similarly, we can investigate the combined effect of the two highest stages by taking the third stage units as the PSUs for the computation of sampling errors. In the extreme case, all stages except the ultimate (as well as other complexities of the design) are assumed absent to estimate the overall design effect, as done in Sec. 5.2.

Consider, for instance, a design with three area stages (say counties, communes, villages) followed by sampling of households. By regarding this as a single stage sample of households, we estimate the variance (v_o) of a SRS of the same size. Variance (v_1) computed by considering villages as the PSUs gives the relative increase (v_1/v_o) due to the clustering of sample households into villages. Similarly, computation with communes treated as the PSUs gives variance v_2, and with counties as the PSUs gives variance v_3 of the actual design. The ratio v_2/v_1 is the effect of clustering sample villages into communes, and v_3/v_2 is the effect of the clustering of the communes into counties.

A similar procedure can be used to investigate the effectiveness of stratification in improving the efficiency of the design. Variances computed for the actual stratified design can be compared with those computed by disregarding stratification. Indeed, the two ideas can be applied in combination: the effect of stratification at various levels (stages) can be investigated by comparing the computations, with and without stratification taken into account, for the sample with the actual number of stages, or with one or more of the highest stages removed. The linearisation method of Section 2.2 is usually the most suited for such analysis. Examples of application of these procedures are given in Illustration 5C.

ILLUSTRATION 5C SOME EXAMPLES OF THE EFFECTS OF SAMPLING STAGES AND STRATIFICATION.

<u>The Effect of Sampling Stages</u>

Table 5C.(1) provides some numerical results on variance components from three samples used in national fertility surveys in Thailand, Colombia and Nepal (Verma et al, 1980).

In Thailand (Rural), the sample consisted of a design with four stages. Changwat (provinces), which are large units averaging over half a million in population, served as the PSUs; Amphoe (communes) as the second stage units (SSUs); villages as the third stage units; and households as the ultimate units of sampling. (The total number of units selected at the four stages were 37, 78, 234 and 3240 respectively.) Generally, area units at various stages were selected with systematic PPS after geographical and administrative stratification, and the final sample of households and women was approximately self-weighting. To investigate the effect of the three area stages, four sets of computations were made for each of a large number of variables over the total sample and various subclasses. The variables concerned fertility and associated factors such as marriage and contraception.

In the table, the results shown are averaged over groups of similar variables. For each statistic, the four variances computed were: v_o for an equivalent simple random sample of the same size; v_1 for a design with one area stage (villages as the PSUs); v_2 for a design with two area stages (communes as PSUs and villages as SSUs); and v_3 for the actual design with three area stages. The quantities S1, S2 and S3 shown in the table are square-roots of the ratios v_1/v_o, v_2/v_o and v_3/v_o respectively. Here S1 is the design effect (deft) for the hypothetical design with one area stage (villages as PSUs); S2 stands for deft for the design with two area stages (communes as PSUs), and S3 for deft for the actual design with three area stages. Thus due to the clustering of households and women within villages, the standard error is inflated by the factor S1; it is <u>further</u> inflated by the factor S2/S1 due to the clustering of villages into communes, and by S3/S2 due to the clustering of communes into provinces. The relative contribution of stages varied by nature of the variable; also the increase due to higher stages was generally smaller for subclasses than for the total sample.

The three stage sample of Colombia (Rural) consisted of a relatively small number (35) of rather large PSUs, from each of which a large number of small units (clusters) were selected as the SSUs, and finally an average of only 3.2 women per cluster were selected at the last stage. Results are shown for the same set of variables as for Thailand and Nepal. As a result of the above design, most of the increase in standard error over SRS comes from the first stage, ie clustering of SSUs into sample PSUs: the overall average of S2/S1 for all variables over the total sample is 1.6, indicating an increase in variance by a factor $(S2/S1)^2$ of over 2.5. The impact was less marked when estimates over subclasses were considered.

The sample for Nepal was more complicated and more heavily clustered due to difficult travel conditions in the country. Only 7 blocks were selected in the urban sector in a single stage. In the rural sector, which comprised 96% of the total sample, 33 districts (PSUs) were selected followed by 2 panchayats per sample district. From 66 sample panchayats, usually only one but sometimes two and occasionally three wards were selected, resulting in a sample of 95 rural wards. The results in Table 5C.(1) were computed by regarding the design as a two stage sample, with 40 PSUs (7 urban blocks and 33 rural districts) and 102 SSUs (7 urban blocks and 95 rural wards). There was no subsampling within SSUs, from which a total of nearly 6000 women were interviewed. Hence the sample consisted of relatively large compact clusters, themselves clustered into a small number of PSUs. Consequently there is a large effect of clustering on the sampling error, with $v_1/v_o = S_1^2 = 3.0$ due to clustering of women within SSUs, and further increase by a factor $(S2/S1)^2 = (1.36)^2 = 1.85$ due to clustering of the SSUs. Again, these effects were greatly reduced for estimates for subclasses such as particular age groups, and especially for small subclasses such as women with higher education.

TABLE 5C.(1). THE EFFECT OF THE NUMBER OF SAMPLING STAGES ON THE MAGNITUDE OF THE SAMPLING ERROR.
(Source: Verma et al, 1980)

Change in sampling error due to area stages—for the total sample and for selected subclasses[1]

	Thailand—rural						Colombia—rural			Nepal—total		
	S3[2]	S2	S1	S2/S1	S3/S2	S3/S1	S2[2]	S1	S2/S1	S2[2]	S1	S2/S1
BY VARIABLE GROUP												
Total sample												
Nuptiality	1·26	1·08	1·10	0·98	1·17	1·15	1·40	1·06	1·32	1·90	1·50	1·27
Fertility	1·34	1·28	1·23	1·04	1·05	1·09	1·33	1·02	1·31	1·92	1·36	1·40
Preferences	1·35	1·34	1·19	1·12	1·01	1·13	1·26	1·04	1·21	2·76	1·97	1·40
Knowledge	2·54	2·47	1·91	1·30	1·03	1·33	3·05	1·31	2·33	3·21	2·43	1·32
Use	2·30	2·10	1·78	1·18	1·09	1·29	2·24	1·12	2·00	2·23	1·72	1·29[3]
All variables	1·66	1·55	1·39	1·12	1·07	1·20	1·75	1·09	1·61	2·32	1·70	1·36
Subclass												
'Age 25–34'												
Nuptiality	1·22	1·03	1·07	0·96	1·19	1·14	1·05	1·05	1·00	1·46	1·24	1·18
Fertility	1·16	1·17	1·19	0·99	0·99	0·98	1·08	1·02	1·06	1·41	1·18	1·20
Preferences	1·16	1·15	1·13	1·02	1·01	1·03	1·10	1·02	1·08	1·94	1·59	1·22
Knowledge	1·5	1·44	1·25	1·15	1·11	1·28	1·85	1·17	1·58	2·26	1·65	1·37
Use	1·50	1·45	1·32	1·10	1·09	1·19	1·65	1·08	1·53	1·76	1·31	1·34[3]
All variables	1·31	1·23	1·19	1·03	1·06	1·10	1·31	1·06	1·24	1·68	1·35	1·24
Subclass												
"High education"												
Nuptiality	0·86	0·89	0·92	0·97	0·96	0·93	1·02	1·01	1·01	1·20	1·12	1·08
Fertility	1·03	1·02	1·03	0·99	1·01	1·00	1·01	1·00	1·01	1·01	0·96	1·05
Preferences	0·97	0·99	1·05	0·95	0·98	0·92	0·95	1·02	0·93	1·17	1·12	1·05
Knowledge	0·98	1·01	1·00	1·01	0·97	0·98	0·90	1·04	0·87	1·17	1·15	1·02
Use	1·01	1·14	1·09	1·09	0·93	1·02	0·98	1·05	0·93	1·29	1·01	1·28[3]
All variables	0·94	0·97	0·97	1·00	0·97	1·00	0·98	1·02	0·96	1·11	1·04	1·06

[1] S_j = *deft* computed by considering only the lowest *j* area-stages. [2] *deft* for the actual (two or three area stage) design. [3] Based on a single variable in Nepal. [4] See Table 2 for description of variables

TABLE 5C.(2). THE EFFECT OF STRATIFICATION ON THE MAGNITUDE OF THE SAMPLING ERROR.
(Source: Verma et al, 1980)

Change in sampling errors due to stratification—for the total sample and for selected subclasses†

	Thailand—rural				*Colombia rural*			*Nepal*			*Indonesia*		*Bangladesh*		*Guyana*	
	S3‡	S3/U3	S2/U2	S1/U1	S2‡	S2/U2	S1/U1	S2‡	S2/U2	S1/U1	S1‡	S1/U1	S1‡	S1/U1	S1‡	S1/U1
TOTAL SAMPLE																
Nuptiality	1·26	0·82	0·69	0·87	1·40	0·77	0·95	1·90	0·95	0·94	1·45	0·92	1·22	0·95	1·16	0·85
Fertility	1·34	0·75	0·83	0·92	1·33	0·85	0·97	1·92	1·00	0·91	1·41	0·87	1·12	0·92	1·02	0·94
Preferences	1·35	0·76	0·85	0·92	1·26	0·90	0·98	2·76	0·87	0·83	1·55	0·92	1·20	0·89	1·13	0·96
Knowledge	2·54	0·82	0·89	0·89	3·05	0·93	0·94	3·21	0·91	0·89	2·44	0·79	1·65	0·78	1·51	0·97
Use	2·30	0·78	0·82	0·89	2·24	0·94	0·95	2·23	1·09	0·97	1·70	0·79	1·31	0·88	1·24	0·93
All variables	1·66	0·78	0·81	0·90	1·75	0·88	0·96	2·32	0·94	0·89	1·62	0·85	1·26	0·89	1·17	0·93
SUBCLASS "AGE 25–34"		(M_s = 0·30)				(M_s = 0·27)			(M_s = 0·34)		(M_s = 0·32)		(M_s = 0·29)		(M_s = 0·34)	
Nuptiality	1·22	0·91	0·84	0·95	1·05	0·78	0·98	1·46	1·02	0·98	1·26	0·91	1·09	0·96	1·10	0·94
Fertility	1·16	0·74	0·84	0·94	1·08	0·84	0·99	1·41	0·92	0·92	1·30	0·91	1·06	0·93	1·05	0·95
Preferences	1·16	0·80	0·87	0·95	1·10	0·92	0·99	1·94	0·84	0·86	1·31	0·93	1·09	0·92	1·07	0·96
Knowledge	1·59	0·86	0·86	0·91	1·85	0·96	1·05	2·26	0·99	0·91	1·76	0·81	1·38	0·87	1·16	1·01
Use	1·58	0·78	0·82	0·91	1·65	0·98	1·01	1·76	1·08	0·97	1·43	0·87	1·17	0·89	1·07	0·91
All variables	1·31	0·80	0·84	0·93	1·31	0·90	1·00	1·68	0·93	0·91	1·37	0·89	1·13	0·92	1·08	0·95
SUBCLASS "HIGH EDUCATION"		(M_s = 0·04)				(M_s = 0·06)			(M_s = 0·06)		(M_s = 0·32)		(M_s = 0·20)		(M_s = 0·16)	
Nuptiality	0·86	0·97	1·00	1·01	1·02	0·94	1·00	1·20	1·03	1·01	1·29	0·95	1·08	0·97	1·06	0·97
Fertility	1·03	1·00	0·99	1·00	1·01	1·05	1·00	1·01	0·92	0·85	1·27	0·94	1·13	1·01	1·06	1·00
Preferences	0·97	0·96	1·00	1·01	0·95	1·05	1·00	1·17	0·93	0·92	1·31	0·93	1·19	0·97	1·06	1·02
Knowledge	0·98	0·97	1·01	1·00	0·90	0·77	1·00	1·17	0·86	0·90	1·72	0·84	1·33	0·96	1·17	0·97
Use	1·01	0·89	0·97	0·98	0·98	0·79	1·00	1·29	0·86	0·79	1·47	0·90	1·08	0·88	1·06	0·99
All variables	0·94	0·96	0·99	1·00	0·98	0·92	1·00	1·11	0·92	0·89	1·51	0·91	1·14	0·95	1·07	0·99

† S_j = *deft* computed by considering only the lowest j area-stages, retaining actual stratification of PSUs; U_j = same, but disregarding stratification
‡ *deft* for the actual (one, two or three area-stage) stratified design.

The Effectiveness of Stratification.

Stratification is a more powerful instrument for controlling variance in multistage samples than in random samples of elements. In addition to the control and flexibility in design offered by stratification, the gains in precision tend to be greater when sampling clusters. Also the more marked the effect of clustering, often the more marked is the proportionate gain due to stratification. To estimate the effect of stratification, a procedure similar to the above can be followed: the actual sampling error can be compared with the (generally increased) error which would be obtained if the units were not stratified. The latter is computed by simply ignoring the actual stratification of the design. Clearly, the linearisation method (Chapter 2) is the appropriate one for this purpose. It is the stratification at the first stage ('primary stratification') which is the most important in many designs. Table 5C.(2) shows some results from a number of national fertility surveys (Verma et al, 1980). To illustrate the effect of stratification for designs with different numbers of stages, and for the total sample and subclasses, the table shows the ratio of two quantities

S_j the standard error computed with actual stratification for a design with j area stages; and

U_j the standard error computed by ignoring the primary stratification in the design with j area stages.

Cases where j shown is smaller than the actual number of area stages in the design mean that the lowest j area stages only were considered in the computations. 'Primary stratification' always refers to the stratification of the actual PSUs. Thus in a design with 3 area stages, S3/U3 is the proportionate reduction in standard error obtained by the stratification of the PSUs; S2/U2 is the reduction in a hypothetical design with the lowest two area stages only; and similarly S1/U1 is the reduction in a design with a single area stage where in computing S1 the first two stages are ignored, and in computing U1 the stratification of the actual PSUs is also ignored. Generally the results show increasing effect of stratification with increasing number of sampling stages. These effects are less marked for subclasses of the type shown than for results over the total sample.

5.6 COMPONENTS OF VARIANCE BY SAMPLING STAGE

The information in the preceding illustration can also be presented in the more conventional form of additive components of variance attributed to the various stages of sampling. Such an analysis of variance into components is the topic of this section. As noted in Section 5.1, the decomposition of overall variance into components involves complex procedures, which need to be more specific to detailed features of the sample design. Often the results obtained are numerically unstable. For these reasons, the treatment in this section is selective and relatively brief.

For specificity, consider the estimation of a ratio from a general three stage stratified design, with sampling without replacement at all stages. The objective is to decompose the overall variance V into three components by stage and estimate these components from the sample.

$$V = V_a + V_b + V_c = \sum_h (V_{ha} + V_{hb} + V_{hc}) = \sum_h V_h. \qquad (5.13)$$

V_a is the between-PSU component, V_b is the component between SSUs within PSUs and V_c is the component between ultimate units within SSUs. Each component may be decomposed into subcomponents by stratum as done in the second part of (5.13). Subscripts h, i, j and k are used to denote primary strata, PSUs, SSUs and USUs respectively. Thus with w as the sample weights, the ratio of two aggregates y and x is estimated as

$$r = \frac{y}{x} = \frac{\sum w_{hijk} \cdot y_{hijk}}{\sum w_{hijk} \cdot x_{hijk}} \tag{5.14}$$

where the summation is over all k, j, i and h in turn. As described in Section 2.2, var(r) can be expressed in terms of an auxiliary variable z defined as follows:

$$\begin{gathered} z_{hijk} = w_{hijk} \cdot (y_{hijk} - r . x_{hijk}) / x \,, \\ z_{hij} = \sum_k z_{hijk}; \quad z_{hi} = \sum_j z_{hij}; \quad z_h = \sum_i z_{hi}. \end{gathered} \tag{5.15}$$

The following treatment is based on Kish (1965; section 8.65). Variations in sampling rates in different parts of the sample complicate the estimation formulae greatly. To begin with we assume that within each stratum the sampling rates are uniform at each stage (f_{ha}, f_{hb}, f_{hc}), or that values appropriately averaged to the stratum level can be used.

Define the following sample quantities

$$v_{ha} = (1-f_{ha}) \cdot \left[\frac{a_h}{a_h - 1} \cdot \left(\sum_i z_{hi}^2 - z_h^2 / a_h \right) \right]; \tag{5.16}$$

$$v_{hb} = (1-f_{hb}) \cdot \sum_i \left[\frac{b_{hi}}{b_{hi} - 1} \cdot \left(\sum_j z_{hij}^2 - z_{hi}^2 / b_{hi} \right) \right]; \tag{5.17}$$

$$v_{hc} = (1-f_{hc}) \cdot \sum_i \sum_j \left[\frac{c_{hij}}{c_{hij} - 1} \cdot \left(\sum_k z_{hijk}^2 - z_{hij}^2 / c_{hij} \right) \right]; \tag{5.18}$$

$$v_h = v_{ha} + v_{hb} + v_{hc} \; ; \quad v = \sum_h v_h . \tag{5.19}$$

In the above, a, b and c refer to the number of units selected at the first, second and third stages respectively. It can be shown that v estimates variance V of the ratio r, and the relationship between the above sample quantities and the required variance components in (5.13) is as follows.

VARIANCE COMPONENT:	*ESTIMATED BY*
(i) V_{hc}:	v_{hc}
(ii) V_{hb}:	$v_{hb} - (1-f_{hb}).v_{hc}$
(iii) V_{ha}:	$v_{ha} - (1-f_{ha}).(v_{hb} - f_{hb}.v_{hc})$
(iv) $V_h = V_{ha}+V_{hb}+V_{hc}$:	$v_{ha} + f_{ha}.v_{hb} + f_{ha}.f_{hb}.v_{hc}$
(v) $V = \sum_h V_h$:	$v.$

The first three may be summed over h to obtain estimates V_a, V_b, and V_c. Note that (iv) is a more precise expression than the one given in equation 2.13, Section 2.2. The former takes into account the without-replacement character of sampling, while the latter assumes sampling with replacement at all stages except the last. It can be seen that the procedure in Section 2.2 amounts to estimating V_b by:

$$(iv') \quad V_h = (1-f_h).\left[\frac{a_h}{a_h-1}.\left(\sum_i z_{hi}^2 - z_h^2/a_h\right)\right] = \left[\frac{1-f_h}{1-f_{ha}}\right].v_{ha},$$

where f_h is the uniform overall sampling rate in the strata:

$$f_h = f_{ha}.f_{hb}.f_{hc}$$

In most practical situations in national surveys, the sampling rates are small and (iv) and (iv)' do not differ significantly, the latter usually providing a slight overestimation. The great difference between the two is that (iv)' is much simpler as it only involves quantities aggregated to the PSU level, and complexities of subsampling within PSUs do not appear in the computational formulae. It is for this reason that (iv)' forms the basis of a general method applicable to diverse designs. (See also the particular case discussed in Section 2.5.)

Instability of the sample estimates of various quantities above is a real practical problem in decomposition of variances into stages. If at each stage, the sampling rate is uniform across all strata (or a suitable average value can be used), then one may aggregate quantities like v_{hj} across strata before using them to estimate the variance components. For instance, in place of (iii) we have

$$(iii)' \quad V_a = \sum_h v_{ha} - (1-f_a).\left(\sum_h v_{hb} - f_b.\sum_h v_{hc}\right) \tag{5.20}$$

Another point worth mentioning is that in most practical designs area units are selected with PPS, so that the assumption of uniform sampling rates within or across strata is not valid. However, it is reasonable to take the following uniform value in such a situation:

effective sampling rate (at a given stage, in a particular stratum if applicable)

$$= \frac{\Sigma(\textit{measure of size of the units selected with pps})}{\textit{total measure of size of all units in the population}}$$

6

DATA REDUCTION AND MODELLING

6.1 OBJECTIVES

The typically large number of estimates produced in national household surveys raised two basic issues: how to be economical and selective in undertaking the sampling error computations for the required statistics; and how to summarise, analyse and make the best use of the information resulting from the computations. The first issue was considered in Chapter 4 and the second is the topic of Chapter 7. This chapter is concerned with the relationship between the two, developing further some ideas introduced in Chapter 5. The above mentioned two issues are closely related because both are served by exploration of the pattern of variation of sampling errors for diverse statistics. This exploration requires data reduction and modelling of the relationships between measures of sampling error over different statistics computed over diverse population bases. By 'data reduction' we mean removing superficial details and variability from numeric data, appropriately amalgamating them, and computing various measures summarising their essential features, so as to identify patterns and relationships that exist in the data. 'Modelling' means describing and generalizing these patterns and relationships in concise, possibly analytical, forms. Data reduction and modelling have several inter-related objectives in the context of sampling error analysis.

[1] Limiting the Volume of Computations.

As discussed in Chapter 4, generally it is not possible (nor often necessary or useful) to compute sampling errors for each and every of the hundreds or thousands of estimates produced in a survey. However, a proper selection can be made only on the basis of an investigation of patterns of similarities and differences in sampling errors for diverse statistics. For instance, groups of similar variables and similar subpopulations need to be identified so that computations, while aiming to cover the diverse groups to the maximum extent possible, can be limited within each group without too much loss in the information generated.

[2] Summarisation

Even when sampling errors for a large number of estimates can be computed, it is not possible to present them all in survey reports. Rather than arbitrarily taking a few statistics for publication, it is much better to analyse the data and extract more concise measures describing their essential features for publication. Apart from reduction in volume of the data to be published, there are also more positive reasons for averaging or summarising the information. One important consideration is that sampling error estimates from survey data are themselves subject to variability. This variability can be particularly large if the computations are based on samples with a small number of primary units (Section 4.4.5). In fact, it is often preferable to use results appropriately averaged over a number of computations, than to rely on the precision of individual computations.

Summarisation is often desirable from the users' point of view as well. Masses of figures are less useful than simple concise presentations from which the required information can be extracted more easily, even if there is some loss of information in the process. While one should aim at providing the user of survey results with all the required information on sampling errors, it is essential to do so in a way that is convenient for the user and that does not obscure the substantive results of the survey, which are after all the primary interest.

Another consideration is that information on sampling errors is not required with the same degree of precision as the information on substantive estimates from the survey. Indeed, too much should not be made of the precise limits of the confidence intervals. For many purposes it is sufficient to have approximate information on the magnitude of the error; more precise information is relevant only where it affects the interpretation of the survey results and the conclusions which are drawn from them. As emphasised in Chapter 1, it should be remembered that sampling error is only one component of the total error which affects the survey results, and that often very limited information is available on the other components.

[3] Extrapolation

Ideally the user of survey results should be able to obtain at least approximate values of the standard error for all estimates derived from the survey, including individual cells in detailed cross-tabulation of the survey results and differences and distributions across cells. Since actual computations cannot be made for all these estimates, it is necessary to establish some means of extrapolation of errors from computations actually made to estimates for which errors have not been computed. Several types of extrapolations may be involved: across diverse subclasses of the sample for a given variable or statistic; across different types of statistics (such as totals and means for the same variable); across different substantive variables; and even across different surveys.

1. Subclasses and differences. The number of subclasses and especially subclass differences of interest in a survey may be extremely large. Hence one of the most important requirements is to be able to extrapolate computed sampling errors (for each variable or statistic) from the total sample to subclasses, across subclasses, and from subclasses to subclass differences. Ideally one would want to be able to generate sampling error for any subclass and any comparison of subclasses in the sample. The pattern may differ by subclass type, such as classes defined in terms of individual characteristics which are distributed over the sample areas, and geographic domains or other aggregates of sample areas.

2. Types of statistics. Extrapolations may also be needed from one type of statistic to another for a given variable: for example from proportions to estimates of total counts; from the more easily estimated sampling error for the

mean to that for the median; or from simple statistics like differences of means to more complex statistics like regression coefficients.

3. Variables. Extrapolation across substantive variables is usually more difficult. Though less common, it is sometimes useful and necessary. It is more difficult because different variables often have major differences in their sampling errors. Such extrapolations are less commonly required because in many surveys the actual number of important variables of interest is often not large - at least in comparison with the numerous subclasses and subclass differences (Section 4.2.2). Nevertheless, many surveys involve groups of similar variables, and it can be useful to average and/or extrapolate sampling error results within such groups.

4. Surveys or survey rounds. Finally there is the requirement of extrapolation across surveys. In a continuing or multi-round survey, the pattern of sampling error results is usually quite stable, and it is possible as well as desirable to pool together and extrapolate sampling error information across survey rounds. Generally, however, extrapolation across surveys can be more difficult than within surveys because of differences between survey conditions, designs, timing, population covered, etc. Nevertheless, such extrapolation is necessary when errors for a survey cannot be computed for some reason; in any case, it is unavoidable for the design of future surveys.

[4] Sample Design and Evaluation

Apart from indicating the reliability of existing survey estimates, an equally important objective of sampling error information is to evaluate how a particular design has fared and to provide data for the design of future surveys. In continuing survey programmes, redesign and improvement of existing samples is also a major concern. In redesign work many of the basic conditions and objectives of the survey often remain unchanged, and the focus is on identifying any major imbalances (inefficiencies) which may be present in the existing design. For well-established surveys, redesign may take the form of fine tuning of the existing design, which requires relatively precise information on sampling and other errors (as well as on cost and operational aspects of the survey). For these purposes, it is necessary to explore patterns of variation of sampling errors as related to important features of sample structure such as clustering, stratification, sample size and allocation, and estimation procedures.

The identification of the relationships of sampling error to the sample structure also helps in meeting the other objectives noted above, namely the objectives of reducing the volume of computations necessary, of analysis and summarisation, and of extrapolating or imputing the information from one situation to another. For the sample design objective it is also important, as noted in Chapter 5, to be able to decompose the total sampling error in multistage designs into components by sampling stages and other aspects of design and estimation procedure; generally this requirement is less important for other objectives.

Relationship Between the Objectives

While there can be differences in emphasis, the various objectives noted above have many common features and requirements. All are helped by an improved understanding of the factors affecting the magnitude of the sampling error, and identification of 'portable' measures which behave in a stable or regular way from one situation to another. For instance, data reduction or summarisation cannot simply be a blind exercise in empirical curve fitting. Rather, it is greatly helped by an understanding of the underlying patterns and relationships. Semi-empirical approaches are usually the best for this purpose: choice of models guided by theory, but with flexibility in the choice of parameters from empirical data, based on actual computations. The same patterns and relationships are also the basis of

extrapolations from one set of samples, variables and subclasses, etc, to another. Even more in-depth information on the pattern of sampling errors and factors affecting their magnitude is required for sample design.
For various purposes it is necessary to combine somehow the results from computations from different variables, subclasses and samples on the basis of which patterns of variation can be established more clearly. However, it is important to recognise that, while smoothing, pooling and extrapolation of computed sampling errors is necessary, there are risks involved in doing that. Excessive or careless application of these procedures can hide real variations, distort the results and mislead the user. The only guarantee against this is to base extrapolation and smoothing on actual computations covering many variables and subclasses of different types, and always to check how well the smoothed or modelled results fit the actual computations. In the following sections, sampling error models of increasing sophistication will be discussed. We begin with the form which in appearance is the simplest: a direct relationship between the magnitude of an estimate and its standard error.

6.2 RELATIONSHIP BETWEEN THE MAGNITUDE OF AN ESTIMATE AND ITS STANDARD ERROR

In many situations it is possible to find a simple (analytical or numerical) relationship between the size of an estimate and its standard error which predicts the actual standard error with acceptable accuracy. The establishment of such a relationship greatly simplifies the task of estimating and presenting sampling errors.
From any estimate already available in the survey report, the reader can obtain an approximate value for its standard error from the relationship between the two expressed in a graphical, tabular or algebraic form. The existence of such relationships is of course conditional on many assumptions about the sample design and nature of the variables involved. But it is very convenient when such relationships do exist.

The possibilities and uses of establishing relationships between estimates and their sampling errors are best demonstrated by considering some illustrations from actual surveys. Four examples are presented below in some detail:

A. Stability of relative errors across similar surveys or survey rounds.

B. Sampling errors of proportions or counts pertaining to different subpopulations in large-scale censuses and surveys.

C. Various approaches to summarising sampling errors for estimated numbers of persons in different categories in labour force and other surveys.

D. Semi-empirical or analytical relationship between the magnitude of an estimate and its sampling error.

It may be noted that, apart from Illustration 6A, the other examples all relate to estimates of proportions or counts, rather than of means or aggregated values of substantive variables. This is because the relationship between an estimate and its standard error is usually more complicated for the latter type of statistics and not amenable to modelling in a simple form.

ILLUSTRATION 6A STABILITY OF RELATIVE ERRORS

By relative error (relative standard error, coefficient of variation) is meant the standard error of an estimate divided by the magnitude of the estimate, often expressed as a percentage. It is found empirically that in certain circumstances relative error is rather stable over time. This applies especially across rounds of a survey with the same or similar content, design, sample size and other aspects of methodology. Stability of the relative error implies that factors other than the size of the estimate itself which affect the standard error remain more or less constant, so that standard error varies directly in proportion to the size of the estimate.

Table 6A.(1) provides an illustration from a series of livestock surveys in Yugoslavia, quoted from Zarchovich (1965). While the estimated total may vary over the years, the relative error tends to be rather stable for a given region and variable. To the extent that relative error can be expected to remain constant, the standard error of an estimate from a new round can be approximated by multiplying the estimate with an averaged value of the relative error from preceding years. Note that in the illustration this procedure has been applied separately within each region.

Table 6A.(2) shows an example from similar data, which in addition examines how stable relative errors are (Zarchovich, 1979). It shows for example that in Quarter 1, averaged over several years, around 45% of the values of relative errors were within ± 5% of the overall average for the category; and 75% were within ± 10% of the average; and 95% within ± 15%. A more comprehensive analysis of variation than that shown in these examples can be carried out by decomposing the total variation into components between quarters, years, items, regions, etc. This can help to identify the best procedures for averaging values for future use.

Stability of relative error is a very simple and convenient model. But at the same time, it involves a number of assumptions which may or may not be valid in particular situations. Even in surveys with the same content and procedures, changes in population characteristics, or in survey conditions, design and sample size, etc, may disturb the stability of the relationship. In any case, continued validity of the assumed relationships should be monitored on the basis of comparisons between new computations and predictions from previous rounds.

TABLE 6A.(1). Relative standard errors in different rounds of a live-stock survey. (Source: Zarchovich, 1965.)

ESTIMATED TOTAL NUMBER OF PIGS AND THE CORRESPONDING SAMPLING ERRORS AS OBTAINED IN A GROUP OF SELECTED DISTRICTS IN THREE SUCCESSIVE YEARS

Serial number	District	Estimated total			Percentage sampling error		
		1957	1958	1959	1957	1958	1959
		.. Thousand			.... Percentage		
1	Kruševac	49	67	81	7 5	6.7	5.4
2	Zaječar	44	47	42	7 0	7.1	6.7
3	Kraljevo	24	30	40	8.2	8.7	8.6
4	Niš	54	66	76	7.4	7.5	7.8
5	Svetozarevo	59	91	102	6.6	6.5	5.3
6	Čačak	38	47	59	5 5	4 4	5.0
7	Pirot	18	21	22	8.4	7.6	7.9
8	Požarevac	117	159	187	6.7	6.7	7.2
9	Prijepolje	6	6	7	13.0	12.0	11.0
10	Prokuplje	26	32	41	5.7	5.9	5.3
11	Smederevo	48	68	80	6.3	5.8	7.6
12	Titovo Užice	28	32	45	6.8	6.8	4.9
13	Novi Pazar	3	4	5	26.5	25.6	25.9

ESTIMATED TOTAL NUMBER OF POULTRY AND THE CORRESPONDING SAMPLING ERRORS AS OBTAINED IN A GROUP OF SELECTED DISTRICTS IN THREE SUCCESSIVE YEARS

Serial number	District	Estimated total			Percentage sampling error		
		1957	1958	1959	1957	1958	1959
		.... Thousand ...			.... Percentage		
1	Kruševac	486	581	444	5.6	4.9	5.0
2	Zaječar	358	367	272	5.1	4.8	4.4
3	Kraljevo	188	226	213	6.1	6 9	6.1
4	Niš	515	580	491	3.8	3.8	3.9
5	Svetozarevo	489	600	478	4.3	4.3	4.7
6	Čačak	352	411	356	4.8	5.1	4.8
7	Pirot	181	197	185	7.4	7.2	7.2
8	Požarevac	691	705	973	5.1	5 2	7.1
9	Prijepolje	47	51	57	6.8	7.2	6 6
10	Prokuplje	254	258	224	5.5	6.0	5.7
11	Smederevo	345	401	347	4.6	4.6	4.6
12	Titovo Užice	233	270	274	5.4	5.2	5.1
13	Novi Pazar	49	49	57	8.8	8 9	8.8

TABLE 6A.(2). An example of stability of relative errors.
(Source: Zarchovich, 1979.)

Relative error in estimated aggregates of number of pigs in Yugoslavia.

	Quarter	1964	1965	1966	1967	1968	1969	average by quarter	overall average
Total									
	1	5.24	5.23	5.10	5.01	5.39	6.11	5.5	
	2	4.22	4.27	4.25	4.34	4.51	-	4.6	
	3	4.44	4.20	4.46	4.33	6.53	6.21	4.8	
	4	5.84	5.53	5.53	5.59	6.46	6.29	5.5	5.1
<2 months									
	1	-	7.14	7.57	7.68	8.21	9.16	8.2	
	2	5.73	6.09	7.18	6.20	7.09	-	6.6	
	3	7.10	6.90	7.06	7.25	9.02	9.18	7.7	
	4	8.76	8.26	8.41	8.92	9.44	9.68	9.0	7.9
2-6 months									
	1	-	6.01	6.66	5.97	6.66	7.37	6.7	
	2	5.07	5.35	4.93	5.23	5.86	-	5.9	
	3	5.38	5.73	5.88	5.54	8.48	7.54	6.5	
	4	6.21	6.07	6.11	5.65	8.48	7.75	6.7	6.5
6-12 months									
	1	-	6.10	5.65	6.25	5.95	6.65	6.0	
	2	5.52	6.34	5.30	5.95	5.34	-	5.5	
	3	5.35	4.95	4.83	4.92	6.57	6.06	5.0	
	4	6.82	6.82	7.11	7.03	7.74	7.56	7.2	5.9
12+ months									
	1	-	6.30	6.92	6.48	7.15	7.45	7.3	
	2	5.46	6.39	6.46	6.93	7.09	-	7.1	
	3	6.26	6.59	6.62	6.93	8.54	8.16	7.4	
	4	6.37	7.15	6.83	7.18	7.75	8.40	7.2.	7.3
Sows									
	1	-	6.25	6.68	6.32	6.91	7.37	7.1	
	2	5.48	6.09	6.08	6.36	6.80	-	6.6	
	3	5.88	6.48	6.39	6.53	8.44	8.04	6.9	
	4	6.48	6.90	6.44	6.44	7.66	8.31	6.8	6.9

Deviation of individual relative errors from the overall average.

	--Quarter 1--		--Quarter 2--	
Interval	no. of errors	cum.%	no. of errors	cum.%
95-105	27	44.3	22	36.7
90-110	19	75.4	14	60.0
85-115	12	95.1	15	85.0
80-120	2	98.4	3	90.0
75-125	-	-	3	95.0
70-130	-	-	3	100.0
65-135	1	100.0	-	-
60-140	-	-	-	-

ILLUSTRATION 6B SAMPLING ERRORS OF PROPORTIONS AND COUNTS

There are many censuses and surveys where the primary interest is in producing estimates of proportions or numbers of individual units which possess certain specified characteristics. For example in a survey of housing, the interest may be primarily in the proportion or numbers of households with access to certain amenities such as running water, electricity, private toilets, etc. Labour force surveys are another good example dealt with in more detail in the next illustration. In such surveys the main focus is usually on proportions and numbers of persons in various categories of the labour force. Similarly in health surveys, the interest may be in estimating proportions or numbers of persons in various categories such as those who experience illness or injury, or receive treatment during a specified period. Such estimates are often required separately for numerous geographical subdivisions and groups in the population. The task of computing and presenting sampling errors is greatly reduced when standard errors can be expressed as a simple function of the size of the estimates concerned. Such relationships can be established for estimates of proportions and counts, but not so readily for means or aggregates of values of substantive variables.

Tables 6B.(1) and (2) present two parts of the relationship between the sampling error and certain other parameters for an estimated proportion or count. The first part, 6B.(1), gives the relationship on the assumption of simple random sampling (SRS); the second part, 6B.(2), gives the design effects (defts) by which standard errors from the first part may be multiplied to obtain the final value of the error for a specific characteristic. In this sense, this illustration represents a more sophisticated (hence more general and accurate) formulation of the relationship than Illustration 6A where differences in deft are not explicitly brought in.

To understand the basis of Table 6B.(1), consider a simple random sample of size n drawn from a population of size N to estimate the proportion p or count N' = p.N of individuals with a certain characteristic. The well known expression for their standard error with SRS is

$$se(p) = \sqrt{(1-f).\frac{p(1-p)}{n}} = \sqrt{\frac{1-f}{f}\,p(1-p)} \;/\; \sqrt{N} \qquad (6.1)$$

$$se(N') = N.se(p) = \sqrt{\frac{1-f}{f}.p(1-p)}.\sqrt{N} = \sqrt{\frac{1-f}{f}.(1-p)}\;.\;\sqrt{N'} \qquad (6.2)$$

$$rse(N') = se(N')/N' = \sqrt{\frac{1-f}{f}.\frac{1-p}{p}} \;/\; \sqrt{N} \qquad (6.3)$$

where se stands for the standard error of an estimate; rse is its value relative to the magnitude of the estimate; and f is the sampling rate = n/N. These equations allows standard errors to be expressed in a very concise form covering the full range of p values encountered.

The data shown relate to a large sample attached to the census of population in the United States (Waksberg et al, 1973), the results for which were required for a number of characteristics (summarised in Table 6B(2)) for numerous geographic subdivisions of the country. Because of the very large number of estimates involved, this form of concise summarisation of the information on sampling errors provides an example of great practical relevance. (A more recent example on the same lines will be given in Illustration 7E in the discussion of modes of presentation of sampling errors in survey reports.)

In the present case, the sample was drawn from the census with a constant rate f = 0.2. With f constant, (6.1) expresses se(p) as a function of p and N. (To remind, p is the proportion of the population with a certain specified characteristic, and N is the total population count in a geographical subdivision of the country.) The magnitude of se(p) is insensitive to the value of p, especially in the middle range on either side of p = 0.5. Hence in the lower panel of Table 6B.(1) it has been considered sufficient to show only a few rows: reasonable values for other values of p may be obtained by interpolation between the rows. Also se(p) is the same for p and its complement q = (1-p). Columns show the error for different population bases N; it is inversely proportional to the square-root of N along any row (fixed p). It is not necessary to go to very large values of the base N, because with large N the corresponding standard error becomes too small to be important.

The upper panel of the table shows se(N') as a function of N (columns) and N' (rows), where N' is the number of persons with a certain characteristic in the total population N. As shown by equation (6.2), for small p = N'/N, standard error se(N') depends only on N', varying approximately proportional to its square-root and remaining practically constant across columns (different population bases, N). The effect of N appears only when p is large, and generally remains small. The important thing is that for an SRS, the same relationships (6.1) and (6.2) are valid independently of the particular characteristic defining p, or the particular base population (such as a geographic area) being considered. The main assumption of the model comes in moving from SRS to actual standard errors through the introduction of defts in Table 6B(2). The value of deft depends on various factors but, on the basis of empirical information in the present case, it depends predominantly on the particular group of characteristics being considered. This assumption is the basis for constructing Table 6B(2). The values shown are actually averaged over a large number of computations for different characteristics in each group and over different geographical domains. (The last column in the table shows the range of values averaged.) Different such tables were in fact constructed for different major regions of the country.

TABLE 6B.(1). Standard errors of estimated counts and proportions assuming simple random sampling
(Source: Waksberg et al, 1973.)

Approximate Standard Error of Estimated Number Based on 20-Per cent Sample

Estimated number[1])	Number of persons in area[2])								
	1,000	10,000	25,000	100,000	250,000	1,000,000	3,000,000	5,000,000	20,000,000
50	15	15	15	15	15	15	15	15	15
100	20	20	20	20	20	20	20	20	20
250	30	30	30	30	30	30	30	30	30
500	30	45	45	45	45	45	45	45	45
1,000	.	60	60	65	65	65	65	65	65
2,500		90	95	100	100	100	100	100	100
5,000		100	130	140	140	140	140	140	140
10,000		.	150	190	200	200	200	200	200
15,000	...	...	150	230	240	240	240	240	240
25,000	...	...	...	270	300	310	310	320	320
50,000	...	...	...	320	400	440	440	440	450
75,000	.	...	...	270	450	520	540	540	540
100,000	...	...	...	.	490	600	620	630	630

[1]) For estimated numbers larger than 100,000, the relative errors are somewhat smaller than for 100,000.
[2]) An area is the smallest complete geographic area to which the estimate under consideration pertains. Thus, the area may be the state, city, county, standard metropolitan statistical area, urbanized area, or the urban or rural portion of the state or county. The rural farm or rural nonfarm persons in the state or county, the Negro persons, etc., do not represent complete areas.

Approximate Standard Error of Estimated Percentage Based on 20-Per cent Sample

Estimated percentage	Base of Percentage						
	500	1,000	2,500	10,000	25,000	100,000	250,000
2 or 98	1.3	0.9	0.6	0.3	0.2	0.1	0.1
5 or 95	2.0	1.4	0.9	0.4	0.3	0.1	0.1
10 or 90	2.7	1.9	1.2	0.6	0.4	0.2	0.1
25 or 75	3.9	2.7	1.7	0.9	0.5	0.3	0.2
50	4.5	3.2	2.0	1.0	0.6	0.3	0.2

TABLE 6B.(2). Defts to be applied to Table 6B.(1) to obtain standard errors for different types of characteristics.

Factor to be Applied to Standard Error

Subject	Sample rate (per cent)	Average factor	Range of factors
Race			
South	20	0.9	0.9 – 1.0
Other regions	20	1.6	1.0 – 1.7
Age	20	0.8	0.8 – 0.9
Household relationship	20	0.5	–
Families and subfamilies[1])	20	0.6	–
Unrelated individuals	20	1.3	1.2 – 1.4
Type of group quarters	20	0.6	0.5 – 0.6
Marital status	20	0.6	0.6 – 0.7
Marital history	5	2.0	1.9 – 2.1
State of birth	20	1.3	1.2 – 1.5
Country of origin	15	1.6	1.5 – 1.8
Spanish origin or descent	5	2.9	2.7 – 3.3
Nativity and parentage	15	1.7	1.4 – 1.9
Mother tongue	15	1.8	1.6 – 2.0
Year moved into present house	15	1.9	1.7 – 2.0
Residence in 1965	15	2.0	1.8 – 2.2
Rural farm-nonfarm residence			
United States, total	20	1.7	1.5 – 2.0
Inside SMSA	20	1.5	–
Outside SMSA	20	1.9	–
School enrollment	15	1.0	0.9 – 1.0
Years of school completed	20	1.0	1.0 – 1.1
Vocational training	5	1.7	1.6 – 1.8
Veteran status	15	0.9	0.9 – 1.0
Disability	5	2.4	2.2 – 2.6
Labor force status	20	0.8	0.7 – 0.8
Unemployed	20	1.1	1.0 – 1.3
Weeks worked in 1969	20	0.8	0.7 – 0.8
Activity 5 years ago	20	0.8	0.7 – 0.8
Place of work	15	1.3	1.2 – 1.3
Means of transportation to work	15	1.3	1.2 – 1.3
Occupation	20	1.1	1.0 – 1.1
Industry	20	1.1	1.0 – 1.1
Class of worker	20	1.1	1.0 – 1.1
Income			
Persons	20	1.0	0.9 – 1.1
Families	20	1.0	1.0 – 1.1
Poverty status			
Persons	20	1.8	1.7 – 2.1
Families	20	1.1	1.0 – 1.2
All other			
20 per cent	20	1.0	–
15 per cent	15	1.2	–
5 per cent	5	2.2	–

ILLUSTRATION 6C ESTIMATES OF COUNTS IN DIFFERENT LABOUR FORCE CATEGORIES.

Many surveys have the objective to estimate the number of households or persons in various categories of the population. This is the case for example with labour force surveys, where most of the statistics of interest take the form of estimates of the number of persons in various economic categories by employment, occupation and industry, etc. Separate estimates are usually required for different geographical areas and diverse demographic and socioeconomic subclasses in the population. This makes it highly desirable to be able to relate standard errors to the size of the estimated counts. The relationship may have to be specified separately for different geographical areas, different classes of the population, different labour force categories, or different periods of the survey. At the same time, it is highly desirable to model the information so as to minimise the number of separate presentations needed. Because of the great practical importance of the issues involved, several examples are presented and discussed below (International Labour Office, 1986). These involve the presentation of sampling errors in concise and abbreviated form to various degrees. This is done on the basis of some underlying model which has not always been made explicit in the examples given. Some insight can be gained by comparing the various forms in Table 6C with the simple model of Illustration 6B.

As a point of reference, Table 6C.(1) shows the crude approach of presenting actual values of standard error for a selection of specific estimates, such as for the number of employed or unemployed persons in a particular age or sex group. While such information may be useful in interpretation of the specific results, its limitation is that it provides no direct information on the pattern of variation of sampling errors, nor on many other subgroups of interest not explicitly included.

TABLE 6C. Standard errors in labour force surveys:
[1] Venezuela.

	Estimated figure	Standard error
Population aged 15+	8399945	66607
In the labour force	4494689	38732
males	3423711	30369
females	1260973	15491
Total employed	4351373	35904
males	3155559	28011
females	1195814	14853
Total unemployed	333316	7956
males	268152	6957
females	65164	3079

Table 6C.(2) goes a little further. By introducing an age and sex classification, the volume of information displayed is considerably increased. In accordance with equation (6.2), the main factor determining se(N'), apart from deft, is the population base N; the dependence on p = N'/N is much less marked. This explains the difference between the 'total' column versus the 'male' or 'female' columns. (The sample base in the latter is roughly half the former, hence their absolute error se(N'), is smaller by a factor of around 0.7.). By contrast, relative error shown in the last three columns is determined by (6.3), apart from deft. It is determined again primarily by N, but this time inversely to its square-root. However dependence on p is more marked than before (because of the different functional relationship; compare the factors under the square-root sign in (6.2) and (6.3)). This explains the big differences in relative errors at the extreme age groups for which p (proportion in the labour force) tends to be much smaller. By showing different population bases, Table 6C.(2) already brings out some regularities in the pattern of sampling errors. However, it suffers from the same basic limitation as Table 6C.(1) in that it provides no information directly on categories not shown. Note also that the table applies only to a particular variable, in this case the population in the labour force; separate such tables would be required for other variables such as the numbers employed or unemployed.

TABLE 6C. Standard errors in labour force surveys:
[2] Finland. Labour force by age and sex. First quarter, 1983.

	Standard error (1000 persons)			Relative standard error (per cent)		
Age group	total	male	female	total	male	female
15-74..........	12	8	8	0.5	0.6	0.7
15-19..........	3	2	2	2.8	4.0	4.0
20-24..........	4	2	3	1.4	1.9	2.2
25-29..........	4	3	3	1.3	1.8	2.0
30-34..........	4	3	3	1.2	1.6	1.7
35-39..........	4	3	3	1.1	1.4	1.6
40-44..........	3	2	2	1.2	1.7	1.7
45-49..........	3	2	2	1.3	1.8	2.0
50-54..........	4	2	3	1.5	2.0	2.3
55-59..........	3	2	2	1.9	2.6	2.8
60-64..........	3	2	2	3.2	4.2	4.6
65-69..........	2	2	1	10.0	13.0	16.0
70-74..........	2	2	1	18.0	21.0	30.0

Tables 6C.(3) onwards follow a different approach. The standard error of an estimate se(N') is shown simply as a function of the size (N') of the estimate. It is assumed that the same relationship to size applies to diverse geographic and other subgroups in the population, independent of the substantive characteristic defining the subgroup. The underlying assumptions become clear by comparing this with the model in Illustration 6B. It is seen from Table 6B.(1) presented earlier that se(N') is practically constant for different population bases (N), except for very large N'/N (generally the first one or two entries in any row). In Table 6C.(3), this variation with N is ignored or averaged away. Furthermore, this table applies to a particular variable (such as size of the labour force) or to a group of similar variables, for which deft may be assumed uniform over different domains or subclasses and incorporated directly into the table. This is the same assumption as in Illustration 6B, except that a separate table of defts is not needed for the reason noted above.

TABLE 6C. Standard errors in labour force surveys:
[3] Singapore. (Based on the 1983 survey results)

Size of Estimate	Standard Error	Size of Estimate	Standard Error
2 000 000	2 758	5 000	308
1 000 000	3 373	2 000	195
500 000	2 754	1 000	138
200 000	1 867	700	115
100 000	1 349	500	97
50 000	964	200	62
20 000	613	100	44

Table 6C.(4) follows the same approach, but shows the relative errors as well. Though relative errors can be useful in the interpretation of specific results, in terms of 'modelling' the patterns of variation, they are in fact less useful than absolute values when estimating proportions or counts (Section 6.4). This is because the relative measure is more sensitive to variations in p among subclasses, variations which have not been properly taken into account in the simplified model on which Table 6C.(4) is based. The table also shows the effect of sample size by comparing quarterly and annual figures, the latter being averaged over four quarterly rounds. The reduction from quarterly to annual averages is by a factor much less than 2 (the square-root of the ratio of sample sizes); this is because of the positive correlation between overlapping quarterly samples.

TABLE 6C. Standard errors in labour force surveys:
[4] Norway.

	Quarterly figures		Yearly averages	
SIZE OF ESTIMATE	STANDARD ERROR Absolute figure	relative value %	STANDARD ERROR Absolute figure	relative value %
5 000	1 400	28.0	900	18.0
7 000	1 700	24.3	1 100	15.7
10 000	2 000	20.0	1 300	13.0
20 000	2 800	14.0	1 800	9.0
30 000	3 400	11.3	2 200	7.3
40 000	4 000	10.0	2 500	6.3
50 000	4 000	8.8	2 800	5.6
60 000	4 800	8.0	3 100	5.2
70 000	5 200	7.4	3 300	4.7
100 000	6 200	6.2	4 000	4.0
200 000	8 600	4.3	5 500	2.8
300 000	10 300	3.4	6 600	2.2
400 000	11 700	2.9	7 400	1.9
500 000	12 800	2.6	8 100	1.6
1000 000	15 900	1.6	10 100	1.0
1700 000	16 000	0.9	10 200	0.6

An important consideration is to improve the accuracy of the model by developing separate versions of the relationship for different types of population subgroups. However, increasing the number of separate versions also has practical disadvantages. Therefore a compromise is required between the two objectives of accuracy and conciseness. In Table 6C.(5), accuracy of the model is improved somewhat by introducing separate versions for males and females. Table 6C.(6) goes much further, by showing the relationship separately for a large number of ethnic, demographic and socioeconomic categories. Presumably, it incorporates differences both in p and deft among the categories.

TABLE 6C. Standard errors in labour force surveys:
[5] Italy. National estimates by sex.

Estimate Y	Men S	Women S	Estimate Y	Men S	Women S
10 000			750 000	12250	12050
20 000	2350	1950	1 000 000	13950	13900
30 000	2850	2400	1 250 000	15450	15500
40 000	3250	2800	1 500 000	16800	17000
50 000	3600	3100	1 750 000	18000	18350
75 000	4300	3800	2 000 000	19150	19600
100 000	4900	4400	3 000 000	23050	24000
150 000	5900	5400	5 000 000	29050	31000
200 000	6700	6200	7 500 000	34950	37950
250 000	7450	6950	10 000 000	39850	43800
300 000	8100	7600	15 000 000	47900	-
400 000	9200	8800	20 000 000	55950	-
500 000	10200	9800	-	-	-

Y: National estimation in absolute figures, from survey results of a specific variable (employment, unemployment, apprentices, students, housewives, etc).

S: Corresponding standard error in absolute figures.

A major issue in the modelling of proportions or counts concerns the identification of groups of the population for which a common form of the relationship between an estimate and its standard error can be used to give results with acceptable accuracy. The basic assumption is that the effects of p and deft are uniform within each group. Here the distinction between cross-classes (defined in terms of characteristics of individuals) and geographical domains is useful (Section 4.2.3). Usually, different domains tend to be similar in the overall p, as each includes a cross-section of individuals with different characteristics. In so far as they form separate design domains, they may differ in the designs used and hence in deft for a given variable. However, it is not uncommon to have similar designs in different geographical domains. Consequently it is often reasonable to use a common form of relationship for diverse geographical domains, as is done in both Illustration 6A and Illustration 6B. By contrast, cross-classes often differ in p, and also in deft which depends on the subclass size (as explained more fully in the next section). This may necessitate the use of different forms of the relationship for different types of subclasses, as done in Table 6C.(7).

TABLE 6C. STANDARD ERRORS IN LABOUR FORCE SURVEYS.
[6] UNITED STATES CURRENT POPULATION SURVEY.

Standard Errors for the CPS Estimates of Monthly Level--Employment and Unemployment
(68 chances out of 100. Numbers in thousands)

Estimated Level	Characteristic																
		Labor Force Data Other Than Unemployment and Agriculture Employment Data														Unemployment	
	Agriculture Employment	Total	White	Black and other	Teenage (16-19)	White Teenage (16-19)	Black and other Teenage (16-19)	Male	Male 20+ or White Male	Black and other Male	White Male 20+	Black and other Male 20+	Female Total or White	Female 20+ Total or White	Black and other Female or Black and other Female 20+	Total or White	Black and other
10	6	5	5	5	5	5	5	4	4	4	4	4	4	4	4	4	5
50 . .	13	10	10	10	10	10	10	9	9	9	9	9	9	9	9	10	11
100 . .	18	14	14	14	14	14	14	13	13	13	13	13	12	12	12	14	15
250 . . .	29	23	23	23	23	23	21	21	21	21	21	21	20	20	19	22	24
500 .	41	32	32	32	32	32	28	30	30	29	30	29	28	28	27	31	33
750 . .	50	39	39	39	38	38	32	37	37	35	36	35	34	34	33	38	40
1,000 . . .	57	45	45	44	44	44	33	42	42	40	42	39	39	39	37	44	46
2,000 . . .	81	64	64	60	60	59	13	59	59	52	59	51	55	55	50	62	63
4,000 .	113	90	90	79	77	74	—	82	82	60	82	55	77	76	61	87	83
6,000	137	109	109	88	84	78	—	99	99	53	98	39	93	92	61	106	93
8,000 . . .	—	125	124	90	84	73	—	113	112	16	111	—	105	105	51	122	—
10,000	—	139	138	87	76	56	—	124	123	—	122	—	116	115	—	135	—
15,000 . . .	—	166	165	36	—	—	—	146	143	—	141	—	137	136	—	162	—
20,000	—	188	185	—	—	—	—	161	157	—	152	—	152	150	—	182	—
30,000	—	219	215	—	—	—	—	177	168	—	159	—	171	165	—	—	—
40,000	—	240	233	—	—	—	—	178	163	—	144	—	177	167	—	—	—
50,000	—	253	242	—	—	—	—	164	137	—	101	—	172	—	—	—	—
60,000 . .	—	260	244	—	—	—	—	131	75	—	—	—	155	—	—	—	—
70,000 .	—	260	238	—	—	—	—	49	—	—	—	—	—	—	—	—	—
80,000	—	254	224	—	—	—	—	—	—	—	—	—	—	—	—	—	—
100,000	—	221	164	—	—	—	—	—	—	—	—	—	—	—	—	—	—
120,000 . . .	—	143	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

The above examples are confined to estimates of proportions or counts. In Table 6C.(7) an attempt has been made to extend this in a very approximate manner to estimates of values, such as aggregated or average hours worked, or average or median duration of employment. The underlying model in this extension is to correct for differences in coefficients of variation (and if necessary defts) between different substantive variables. This extension is not always straightforward and involves additional assumptions and approximations.

TABLE 6C. Standard errors in labour force surveys:
[7] Australia.

Size of estimate	standard error	% of estimate
4 500	970	21.6
5 000	1000	20.0
6 000	1100	18.0
10 000	1400	14.0
20 000	2000	10.0
50 000	2900	5.8
100 000	3900	3.9
200 000	5100	2.6
300 000	6000	2.0
500 000	7200	1.4
1000 000	9100	0.9
2000 000	11000	0.6
5000 000	15000	0.3

Note. The relative standard error of estimates of aggregate hours worked, average hours worked, average duration of unemployment, and medium duration of unemployment are obtained by first finding the relative standard error of the estimate of the total number of persons contributing to that estimate, and then multiplying the figure so obtained by the following factors:

aggregate hours worked = 1.2;
average hours worked = 0.5;
average duration of unemployment = 1.5;
median duration of unemployment = 2.0.

ILLUSTRATION 6D RELATIONSHIP BETWEEN AN ESTIMATE AND ITS ERROR IN A FUNCTIONAL FORM

The empirical task of determining the relationship between the magnitude of an estimate and its sampling error is greatly facilitated if some analytical form of general applicability can be established for this relationship. An example is provided by the following expression used extensively in modelling of sampling errors for estimates of counts (and of proportions using a similar procedure) in the US Current Population Survey:

$$V_x^2 = \frac{var(x)}{x^2} = a + b/x$$

where X is the total number of individuals in a subclass possessing a certain specified characteristic; V_X^2 is its relvariance, ie the square of the relative standard error; and a and b are parameters estimated from actual computations by an iterative least squares procedure. In principle the procedure is applied separately to appropriate groupings of subclasses and items. Once established, it can be used to estimate the variance of other items and subclasses in the group. Basically, the method is applicable to estimates of proportions and counts of the population having a specified characteristic; variance of estimates based on values reported for sample units do not lend themselves to modelling in this way (United States 1978 , Chapter VIII). It may be pointed out that the above model for relvariance implicitly assumes that deft is constant for statistics in any set to which the same model is applied. This follows from the observation that for a proportion p = X/N, using

$$var(p) = deft.\left[\frac{p.(1-p)}{n}\right]^{1/2}$$

the relvariance may be rewritten as

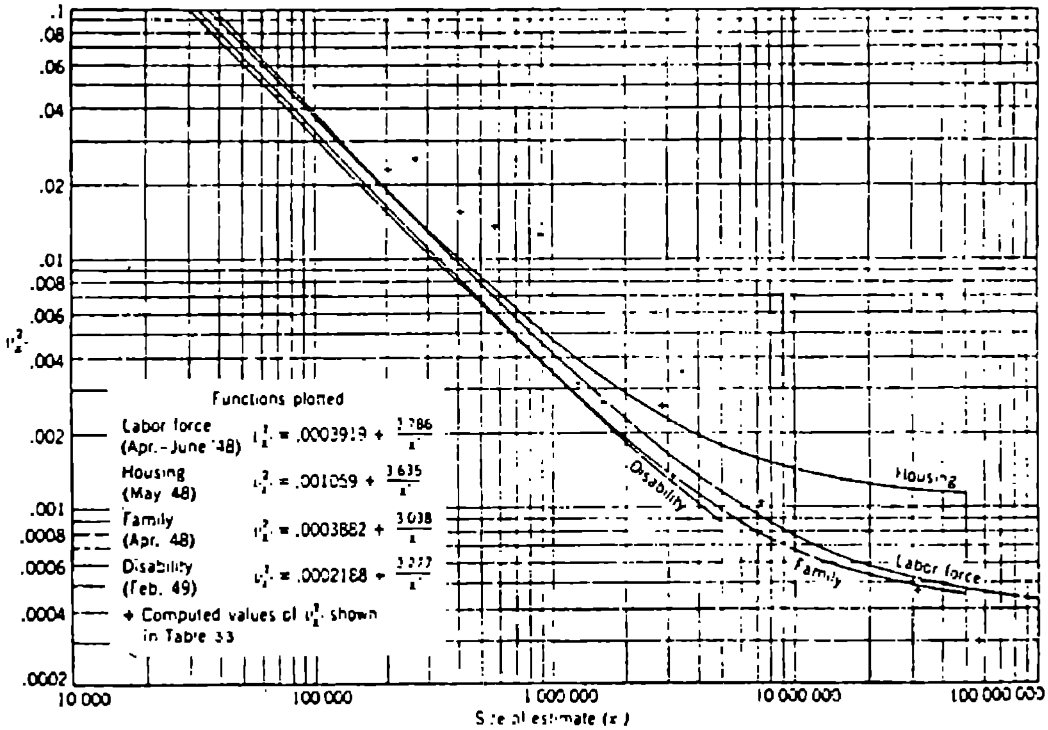
$$V_x^2 = [se(p)/p]^2 = -[\frac{deft^2}{n}] + [N.\frac{deft^2}{n}]/x$$

which with deft constant takes the form (6.4).

The determination of the appropriate grouping prior to estimating parameters a and b is important. It is desirable that the items included in a group have similar design effects. It may be useful to begin with provisional grouping based on judgement and past experience. Scatter plots of V_X^2 versus 1/X could then be examined to identify homogeneous groups. Table 6D.(1) provides an illustration of an application of the model.
The same model is applied to smooth and extrapolate sampling error results in the US Health Interview Survey (Bean, 1970). Table 6D.(2) shows an example of fitted results using the model. 'Type A' data in the graphs refer to prevalence and incidence data collected with a recall period of 12 months, and 'Type B' to certain incidence data collected with a recall period of 2 weeks. 'Medium range' means that values (of incidence etc) for an individual are in the range 0-5 (in contrast to other ranges not shown: narrow range with values 0 or 1, occasionally 2; and wide range with values above 5). This provides a concrete illustration of how the estimates may be divided into groups for fitting the model.

Table 6D.(3) provides an indication of the goodness of fit: it compares the actually computed and modelled value of relvariance of estimated incidence of acute conditions for a large number of subclasses.

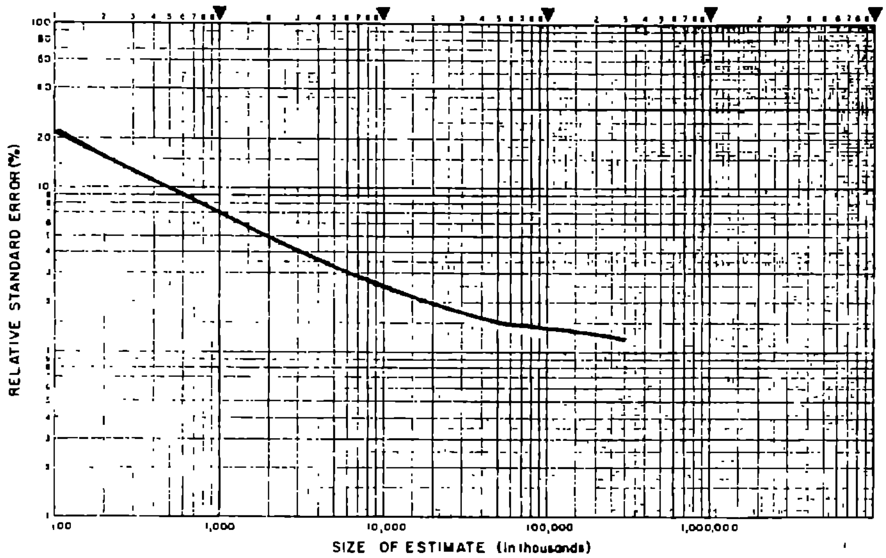
TABLE 6D.(1). An example of functional relationship between an estimate and its rel variance.



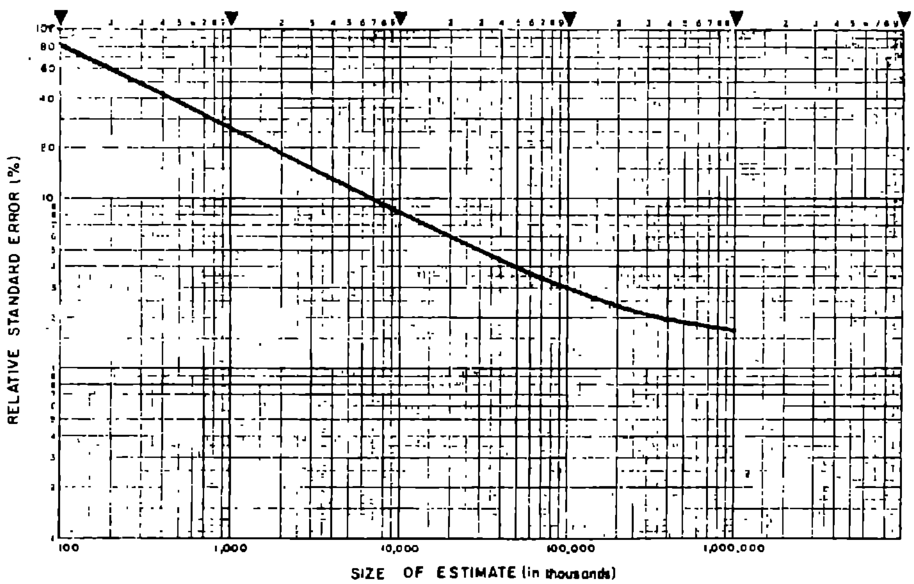
Curves showing relationships between x' and $v_{x'}^2$ for some groups of items. [1]

[1] This figure is reproduced from Morris H. Hansen, William N. Hurwitz and William G. Madow, Vol. 1. p. 575.

TABLE 6D.(2). Examples of functional relationship between an estimate and its relative standard error for two types of data from the same survey.
(US Health Interview Survey; Bean, 1970.)



Smoothed relative standard error curve for aggregate estimates based on 1 year of data collection for Type A data, medium range



Smoothed relative standard error curve for aggregate estimates based on 1 year of data collection for Type B data, medium range

TABLE 6D.(3). Incidence of acute conditions for classes with their actual and smoothed relative variances. United States Health Interview Survey, July 1963 - June 1964. (Source: Bean, 1970.)

CLASS	ESTIMATE (thousands)	RELATIVE VARIANCE Actual	Smoothed
Other respiratory conditions-male, 15-44	1,064	.044017	.057333
Digestive system conditions-male, 45 and over	2,012	.019628	.030473
Digestive system conditions-female, 15-44	4,827	.009058	.012886
injuries-female, 45 and over	5,702	.007866	.010957
All acute conditions-male, retired, 45 and over	6,269	.005896	.009996
Cases of influenza-female, 45 and over	6,408	.006188	.009786
Injuries-female, 15-44	8,402	.005229	.007539
Injuries-male, 25-44	8,441	.006982	.007505
Other respiratory conditions-both sexes, all ages	8,524	.005627	.007435
Digestive system conditions-male, all ages	9,961	.003390	.006408
Contusions and superficial injuries-both sexes, all ages	10,421	.004305	.006139
Other current injuries-both sexes, all ages	12,603	.002849	.005131
Open wounds and lacerations-both sexes, all ages	15,835	.002650	.004148
Fractures, dislocations, sprains, and strains- both sexes, all ages	16,366	.002838	.004024
Digestive system conditions-both sexes, all ages	20,608	.002135	.003261
Upper respiratory conditions-female, 17-44	21,572	.001949	.003129
All other acute conditions -both sexes, all ages	51,941	.001108	.001485
All acute conditions-45-64	52,539	.001071	.001471
Infective and parasitic diseases-both sexes, all ages	55,283	.001491	.001414
All acute conditions - 15-24	55,836	.001053	.001403
Cases of influenza-both sexes, all ages	61,980	.001948	.001296
All acute conditions-under 5	76,083	.000819	.001114
All acute conditions-45-64	79,701	.000747	.001078
All acute conditions-5-14	103,653	.000686	.000902
All acute conditions-currently employed, 17 and over	104,100	.000619	.000899
Upper respiratory conditions-both sexes, all ages	333,797	.000554	.000770
All acute conditions-male	180,182	.000455	.000653
All acute conditions-female	207,175	.000428	.000609
All acute conditions-both sexes, all ages	387,358	.000343	.000473

6.3 PORTABLE MEASURES OF SAMPLING ERROR

The Concept

The magnitude of the standard error of a statistic depends on a variety of factors such as

[1] the nature of the estimate
[2] its units of measurement (scale) and magnitude
[3] variability among elements in the population (population variance)
[4] sample size
[5] the nature and size of sampling units
[6] sample structure; sampling procedures
[7] estimation procedures.

Consequently, the value of the standard error for a particular statistic is specific to the statistic concerned and can, at best, be imputed directly to essentially similar statistics, based on samples of similar size and design drawn from the same population. To relate the standard error of one statistic to that of another, it is necessary to decompose the error into factors which are less specific to any particular statistic, ie factors which are more portable across different statistics. The term 'portability', introduced by Kish, refers to the possibility of carrying over from one survey to another, from one variable to another, or from one population subclass to another, the conclusions drawn regarding some measures of sampling error. A measure portable across a set of statistics implies that its value is the same or similar for all statistics in the set, or that its values can be related in some way. Illustration 6A given earlier provided an example of how relative error may be stable (portable) across similar surveys in particular circumstances. This and other examples discussed previously are special cases of the more general model discussed below. The standard error of (say) a mean may be expressed in several related forms involving measures portable to different degrees:

$$se(\bar{y}) = \bar{y}.rse(\bar{y}) \tag{6.4}$$

$$se(\bar{y}) = deft.sr(\bar{y}) \tag{6.5}$$

$$se(\bar{y}) = deft.s/\sqrt{n} \tag{6.6}$$

$$se(\bar{y}) = deft.(\bar{y}.cv)/\sqrt{n} \tag{6.7}$$

$$deft^2 = D_w^2.[1 + (\bar{b}' - 1).roh] \tag{6.8}$$

Each of these will be explained and discussed in the following.

[1] Relative Standard Error

Relative standard error rse in (6.4) is standard error of an estimate divided by the magnitude of the estimate

$$rse(\bar{y}) = \frac{se(\bar{y})}{\bar{y}}$$

This measure removes the variation in the absolute value of the standard error, se, due to the units of measurement and magnitude of $\bar{y}$, but of course still depends on the other factors determining the magnitude of se. Nevertheless its value may be fairly stable in certain situations, as for example for estimates of a given variable over a given population or subclass from different rounds of a survey with similar design, sample size and methodology. In such situations it may be reasonable to average the relative errors computed from different rounds, and use the average in place of the individual computed values as a more reliable estimate applicable to all the rounds. Or the average computed from past rounds may be used to predict rse for the statistic in future rounds. In this sense the relative measure is more portable across survey rounds, while the actual se varies in proportion to the size of the estimate from one round to another over time. Note that this assumption of portability needs to be empirically established and periodically reconfirmed by fresh computations as new data become available from subsequent rounds. Whatever the measure, there can be diverse sources of variation in sampling error which cannot be controlled or taken fully into account in any model. 'Portability' therefore is a matter of degree; it is a relative rather than an absolute concept.

Note on terminology
The relative standard error is also commonly referred to as the 'coefficient of variation' (cv). However, the latter term is also used for a measure of the relative variability among elements in the population. To avoid confusion, it is preferable to reserve the term cv for use in the last mentioned sense, and use rse for the relative standard error of sample estimates.

[2] Design Effect and Standard Error in an Equivalent Simple Random Sample

In (6.5), standard error for a statistic estimated from a complex sample is factorised into two parts: the design effect, and the standard error which would be obtained with a simple random sample of the same size. This basic decomposition was considered in detail in Section 5.2. Each of the two components can be further decomposed into more portable measures as discussed below.

The measure deft is more portable than the actual standard error, since it does not depend on factors which affect both se and sr in the same way, factors such as units of measurement and the magnitude of the estimate, its variability in the population, and above all, the overall sample size. Consequently, deft is also expected to be portable across a wider range of statistics and situations than relative error, rse, since the latter removes the effect of only some of these factors. For these reasons, deft is one of the most commonly used and useful measures of efficiency of the sampling design.

[3] Population Variance; Standard Deviation

Equation (6.6) isolates in SRS error the effect of the sample size. The concept of population variance (s^2) and its square-root, the standard deviation, was also discussed in Section 5.2. Since s^2 does not depend on the structure or size of the sample, but only on characteristics of elements in the population, it is a very portable measure. Also in most practical samples, it can be estimated well and simply from the sample observations irrespective of the complexity of the design, as detailed in Section 5.2.

[4] Coefficient of Variation

The measure s (or s^2) still depends on the scale of measurement and magnitude of the estimate. This is removed in the coefficient of variation

$$cv = s/\bar{y}$$

The coefficient of variation is more portable than standard deviation. Similarly distributed variables may have similar cv's, irrespective of their actual magnitudes or scales of measurement. However, cv is a useful measure only when the denominator in the above expression is not close to zero, as for example may happen for estimates of differences between subclass means. Also it is more useful for means and ratios than for proportions. For a proportion,

$$cv = s/p = \left[\frac{1-p}{p}\right]^{1/2}$$

may in fact be less portable (more variable with p) than

$$s = [p.(1-p)]^{1/2}$$

which tends to be quite insensitive to p values. Also, unlike s, cv for a proportion lacks symmetry in that it is not the same for a proportion p and its complement (1-p).

Several examples of the portability and uses of the measures s and cv are discussed in Illustration 6E below.

[5] Rate of Homogeneity (roh)

Equation (6.8) decomposes the overall design effect. After removing the effect of increased error due to weighting (D_W) on the lines of Section 5.3, it isolates the effect of the average 'cluster size' (ie the average number of elements selected per sample PSU) to obtain a more portable measure, roh (See Section 6.5).

Unit Variance

Unit variance is another practically useful concept, which combines population variance and design effect. It refers to variance per one sample unit, and applies when the variance or its components vary inversely with the number of units at the corresponding stage in the sample. Unit variance is portable across designs which are identical except for the numbers of sample units.Consider for instance a design in which the conditions of Section 2.2 apply and the variance of a mean is approximately inversely proportional to the number of sample PSUs. Unit variance (variance per PSU) defined as

unit variance = actual variance x no. of PSUs is therefore a portable measure having approximately the same value in designs which have different numbers of PSUs but are identical in other respects. By inverting the above relationship, the effect of the number of sample PSUs on variance can be assessed on the basis of given unit variance, provided that other aspects of the design are fixed. This can be useful in survey design work, and also in imputing results from the total sample to its geographical domains with similar design but differing numbers of PSUs. The concept can be extended to units at different stages.

6.4 DECOMPOSITION OF SRS VARIANCE: POPULATION VARIANCE AND COEFFICIENT OF VARIATION

Portability of the Measures

For a proportion p, the standard deviation s can be computed directly from the estimated proportion (eq. 5.5); there is no need to 'impute' it from other statistics. Furthermore, s is rather insensitive to the exact value of p, especially in a fairly wide range around p = 0.5. As noted earlier, the relative measure cv = s/p is in fact more variable with p, and in this sense less portable across different values of p.

By contrast, for the mean $\bar{y}$ and similar statistics of a substantive variable, s and $cv = s/\bar{y}$ are portable measures, and it is useful to compile information on cv's for different types of variables. In so far as s varies in direct proportion to the scale and size of the estimate, the coefficient of variation can be a stable measure dependent only on the shape of the distribution of variable y in the population. There are many situations in which the above pattern holds. For instance in comparisons over time between different rounds of the same survey, or between geographical domains of the country, or between subclasses in the sample, it is often found that the variation in s reflects differences in y values, so that $s/\bar{y}$ is rather stable across the categories. Of course, for certain types of variables, the relationship between the standard error and the mean can be more complex. For instance Little (1978) notes that for variables which represent cumulative counts (such as cumulative fertility or some other cumulation of events over time), s^2 varies in proportion to the mean, in classes such as age groups closely related to the variable being estimated. In general one may think of a relationship of the form

$$s \;\alpha\; (\bar{y})^{a} \quad ; \quad cv \;\alpha\; (\bar{y})^{a-1}$$

meaning that for variables with <u>a</u> in the range 0.5 to 1.0, cv is a more portable measure than s, but s may be more stable hence preferable for variables with <u>a</u> below 0.5. The value of <u>a</u> has to be empirically determined from computations over different types of variables. Other specific models for cv have for example been noted in United Nations (1989, pp 160-165) in the context of surveys of household income and expenditure.

ILLUSTRATION 6E INFORMATION OF CV's.

One of the important objectives of computing sampling errors in household survey programmes should be to compile information on s and cv values for many variables over different types of subclasses and domains, and to study the pattern of their variation.

[1]. Examples of CV's from Comparative Surveys

Table 6E.(1) provides an example of cv's for means of a few selected variables from fertility surveys in four developing countries. For each survey, estimated values of cv are shown for the total country, the urban sector, one arbitrary region of the country, and for women aged 25-34, (which is a cross-class well distributed over the entire population). For reference, the last two columns of the table show the value of the mean for the total sample and for the age subclass. Several interesting features of the data may be noted.

(1) For a given domain, the most marked differences in cv are by the nature of the variable. Given the relatively strong family size norms, cv's for the variable 'ideal family size' are the smallest, mostly around 0.4-0.45 in the present example; those relating to fertility (children born or alive) are intermediate and show a wider range of variation by country and domain; and cv's are much larger for the remaining variables concerned with post-partum behaviour.

(2) For a given country and variable, urban cv's are generally only slightly higher than rural cv's - this conflicts with the commonly held assumption by survey practitioners that the urban population is generally much more heterogeneous than rural. (Of course such assumptions may be valid for some other types of variables not considered here, and there may be other reasons for oversampling urban areas. However, the present data show that greater urban heterogeneity should not be taken for granted in all cases.)

(3) When the characteristic defining subclasses is closely related to the substantive variable being estimated, it tends to partition the population into more homogeneous groups, thus lowering the cv value within each of the group or subclass defined in terms of categories of the variable. In relation to the age class, this effect is seen most clearly for fertility variables. By using the last two columns, it can be seen that this effect is somewhat stronger in the case of s. in this sense making s less 'portable' than cv.

(4) The above pattern is consistent across countries. This gives confidence in the value and portability of the results across different conditions, at least for a given type of survey. There are notable differences of degree, however. For example, compared with the two Asian countries women in the two African countries are more heterogeneous in terms of fertility, but much more homogeneous in terms of post-partum behaviour. The variable 'ideal family size' clearly illustrates the advantage of cv over s in terms of portability: the former is much more stable over countries and domains, while the latter varies as markedly as the mean values shown in the last two columns.

TABLE 6E.(1). Pattern of cv's in fertility surveys.
(Source: Demographic and Health Surveys, country reports.)

		Total	Urban	a Region	Age 25-34	Estimated Mean Total	Estimated Mean Age 25-34
Ideal Family size	(1)	0.44	0.40	0.55	0.41	3.21	3.16
	(2)	0.41	0.45	0.40	0.36	3.05	2.56
	(3)	0.45	0.41	0.55	-	4.43	-
	(4)	0.43	0.39	0.31	0.42	5.26	5.29
Children ever-born	(1)	0.75	0.75	0.85	0.59	3.39	2.93
	(2)	0.70	0.73	0.71	0.56	3.00	2.19
	(3)	0.88	1.03	-	-	3.66	-
	(4)	0.92	1.02	0.96	0.55	3.16	3.30
No. of living children	(1)	0.74	0.73	0.84	0.58	2.90	2.57
	(2)	0.70	0.71	0.70	0.56	2.83	2.13
	(3)	0.89	1.03	-	-	3.28	-
	(4)	0.94	1.02	0.95	0.57	2.62	2.79
Months child breastfed	(1)	1.26	1.44	1.40	1.17	25.1	24.1
	(2)	1.27	1.81	1.43	1.33	22.7	18.6
	(3)	0.82	0.90	1.00	-	19.4	-
	(4)	0.83	0.97	0.86	0.84	20.3	19.7
Months of post partum amenorrhoea	(1)	2.35	2.65	2.65	2.08	10.9	10.8
	(2)	2.92	3.71	3.13	2.85	7.5	6.5
	(3)	1.39	1.64	1.65	-	10.9	-
	(4)	1.19	1.40	1.27	1.20	14.0	13.7
Months of post partum abstinence	(1)	3.88	4.40	4.64	3.62	5.32	4.81
	(2)	3.28	3.88	3.38	3.59	6.62	4.85
	(3)	2.26	2.47	3.69	-	5.85	-
	(4)	1.29	1.42	1.28	1.37	13.5	12.1

Notes: (1) Indonesia; (2) Sri Lanka; (3) Kenya; (4) Ghana.

[2]. Use in Survey Design

Table 6E.(2) is presented to indicate how information on cv's can be helpful in survey design. The figures actually come from an establishment rather than a household survey (Cyprus, 1990), but still illustrate the point being made. The table shows values of cv's (in percentage terms) for a number of economic sectors (ISCO classification) for a set of important variables which were measured in an annual survey of establishments. Except for employment, the other variables shown are in terms of value per employee. Employment was also measured by selecting establishments with probability proportional to past employment. These are common methods of control in establishment surveys where units differ greatly in size. Consequently the cv values shown are much lower than would be the case if a random sample or population of establishments were considered without controlling for size.

Despite this control, the table shows big variation in cv both by sector and variable, though the pattern is fairly consistent across variables within sectors and vice versa. There are important consequences of the figures for survey design. Averaged over variables, the cv values vary by a factor of 3-4 by sector. Since the required sample size for the same relative precision varies in proportion to the square of cv, this indicates very large differences (perhaps by a factor of up to 10) in sample size requirements for different sectors. Secondly, it is seen that cv's are particularly large for some variables such as direct costs, and especially investment (which is often undertaken in lump sums and infrequently by individual establishments). Consequently these variables are extremely difficult to measure with precision, and may require sample sizes which are too large to be practicable.

TABLE 6E.(2). Coefficients of variation (%) in a population of economic establishments.
(Source: Cyprus, 1990.)

	Variable						
	sales	output	costs	value added	employ-ment	invest-ment	aver. of 2-5
	[1]	[2]	[3]	[4]	[5]	[6]	
Sector:							
31	-	70	78	51	78	246	69
32	-	94	115	61	66	177	84
33	-	10	73	64	76	271	71
34	-	82	87	64	57	169	72
35&39	-	106	117	57	83	253	91
36	-	95	108	59	46	146	77
39	-	113	148	61	70	230	98
61	67	105	131	104	100	217	110
62	27	115	210	111	77	321	128
7112	-	53	65	55	47	411	55
7113	-	72	99	72	97	218	83
7114	-	170	266	123	244	326	201
83	54	55	194	51	47	235	87
63	67	65	84	73	53	430	69
9	97	79	165	69	51	410	91
5	41	108	196	108	89	286	125
5000				50	105		
5100				60	60		
5200				55	70		

Average for employment (variable 5) 80

[3]. Accumulation of Information on CV's

Many statistical packages routinely compute sampling errors without taking into account the sample structure, ie by assuming simple random sampling. While on the whole this represents a serious shortcoming of the packages, it has the useful side-effect that it automatically provides information on cv's. As noted earlier, it is very helpful for survey design to accumulate information on cv's for diverse variables over different subgroups of the survey population. Table 6E.(3) provides some figures (from an unpublished source) for illustration from household income and expenditure surveys in two African countries. These may be useful in other situations to the extent that they are portable.

TABLE 6E.(3). Examples of cv values in income and expenditure surveys

	--Country A-		--Country B--	
	(a)	(b)	(a)	(b)
Income				
from employment	1.1	1.6	1.0	1.7
farming	1.0	1.0	0.9	0.9
other self-employment	1.1	1.7	1.2	1.6
Rent	1.3	1.9	1.3	2.0
Remittance	1.2	1.6	1.6	2.4
Other income	1.5	2.2	1.7	1.7
Total income	0.9	1.1	1.0	1.5
Expenditure and consumption				
Food expenditure	0.9	0.9	0.9	0.9
Other consumer expendi.	1.0	0.9	1.1	1.0
Consumption of own produce				
food	1.1	0.9	0.9	0.8
non-food	1.3	1.8	1.2	1.7
Consumption of goods/services received in kind	0.9	1.3	1.1	2.0
Remittances given	1.3	2.0	1.4	2.4
Total expenditure	0.7	0.7	0.8	0.8

Notes: (a) = mean per household; (b) mean per capita

6.5 DECOMPOSITION OF THE DESIGN EFFECT: THE RATE OF HOMOGENEITY; THE EFFECT OF SAMPLE WEIGHTS

6.5.1 THE BASIC MODEL

Though generally very portable and useful measure, the design effect still depends on other specific features of the design which limit its portability. Therefore it is useful to decompose deft into more portable components to the extent possible.

To begin with, a most useful model is obtained from generalisation of a well-known result for a single stage simple random sample of equal sized clusters. In that design the effect of clustering is to increase variance over SRS of elements by the factor

$$Deft^2 = 1 + (B-1).\rho \tag{6.9}$$

where ρ is the intraclass correlation coefficient measuring the homogeneity of clusters with respect to the variable under consideration, and B is the constant cluster size. It can be shown that the above expression also applies in the presence of random subsampling of elements within clusters; the only change is to replace B by the size of the subsample (b) selected per cluster. Provided that the variation in subsample sizes is not large, a reasonable approximation is provided by using their average value $\bar{b}$ in the expression

$$deft^2 = 1 + (\bar{b}-1).roh \tag{6.10}$$

where roh is a slightly modified version of the intraclass correlation, introduced by Kish (1965; Sec 5.6B). The important point is that roh in (6.10) is essentially the same as the intraclass correlation of complete PSUs in (6.9), at least when the subsampling within PSUs is simple random. This implies that roh is a measure basically independent of the size of the subsamples taken. This makes roh a more portable measure than deft which depends directly on the subsample sizes b as shown by (6.10).

These ideas can be generalised to a multistage design, on the assumption that departures from selfweighting and variations in the 'cluster sizes' are not large. (*Note on terminology*: For simplicity, it is common in the context of a multistage sampling design to use the term 'cluster sizes' to refer to the size of the sample, ie the number of elements, selected per PSU. We will follow this convenient usage, though an expression like 'the size of the primary selection' would be more correct.)

For a given variable, roh depends on the size and nature of the PSUs, and the method of subsampling within the PSUs. With those given, it is essentially independent of the size of the subsamples taken. This makes roh portable across similar designs with different subsampling rates (different b); and in practice even more importantly, portable from the total sample to subclasses which differ from the total sample primarily in the effective 'cluster sizes' involved.

Empirically the dependence of roh on the size of the original PSUs has been noted to be of the form B^{-a}, where a is mostly in the range 0.2 to 0.6 (Hansen et al, 1953, Vol I, pp 306-9). What this means is that it is expected to be lower for larger, more dispersed PSUs, and higher for smaller, more compact units taken as the PSUs. Clustering of sample elements within PSUs - as would result in designs with more than two stages - tends to increase the value of roh above that of the complete clusters (equation 6.9), or equivalently, above that with random subsampling within PSUs. With stratification or systematic subsampling within PSUs, roh may be reduced somewhat.

To summarise, model (6.10) has been developed to provide a reasonable approximation for deft and to separate out the effect of cluster sizes in complex multistage designs, when the overall design is essentially self-weighting and the variation in cluster sizes is not large. Some modifications to the model when these conditions are not met are possible, as noted below. Roh measures the actual intraclass correlation of the complete PSUs to the extent sampling of ultimate units within PSUs is simple random.

A simple example

The following simple example may clarify the relationship of deft and roh to cluster size $\bar{b}$ in (6.10). Suppose that a two stage sample of size around n = 2500 is drawn by selecting 49 clusters and by selecting an average of 51 elements per cluster. Assume also that for a particular variable, $deft^2 = 2$; in other words, the variance of the variable in the clustered sample is twice as large as its variance in an SRS of the same size, meaning that an SRS of n' = 2500/2 = 1250 would have given the same precision. The implied roh value for the variable under this design is:

$$roh = \frac{deft^2 - 1}{\bar{b} - 1} = \frac{1}{50} = 0.02$$

Now suppose that, retaining the same design, the average subsample size is reduced to b = 26. The model expects roh to remain unchanged (portable across the two designs) because the nature of the units and the subsampling procedure has not changed. From (6.10) with roh = 0.02, we get $deft^2 = 1.5$. In the above sense, roh removes the effect of b in deft, and is a more portable measure. However, it should be emphasised that roh is specific to a particular variable, type of sampling units and subsampling procedures. Incidentally, the above example shows that while the SRS variance is doubled by halving the sample size (in terms of the number of ultimate units), variance of the clustered sample (with the same number of higher stage units, but smaller sample takes at the last stage) is increased less rapidly with decreasing sample size, by a factor of only 1.5.

Note that the above fairly dramatic illustration of the effect of changes in sample size is based on the assumption that the estimates being considered are for the total sample, where b and hence deft values are relatively large. However, the effect may be much less marked when results over subclasses are considered. As discussed in more detail in Section 6.5, for subclasses reasonably well distributed over sample clusters, the effective cluster sizes (b) declines in proportion to the subclass size, resulting in smaller deft values for a given roh. With very small b, and hence deft close to 1.0, the impact of overall sample size on variance becomes similar to that in an SRS - variance increasing in direct proportion to the decrease in sample size.

6.5.2 INCORPORATING THE EFFECT OF VARIABLE CLUSTER SIZES

Variability in cluster sizes (ie, in the sizes of the primary selections) is unavoidable in most samples. Often sharp differences are introduced as a result of using different designs in different domains of the sample. The effective cluster sizes for subclasses can be much more variable than those in the total sample. When the cluster sizes vary greatly, it is necessary to modify (6.10) to separate out the increased variance resulting from that variability. A version which has been proposed by several authors is to replace the average cluster size $\bar{b}$ in (6.10) by

$$\bar{b}' = \bar{b}\left[1 + cv^2(b_i)\right] = \frac{\sum_i b_i^2}{\sum_i b_i} \tag{6.11}$$

where $cv(b_i)$ is the coefficient of variation of cluster sizes, b_i. In terms of practical utility of a form like (6.11), the question is always whether it is necessary to introduce complexities in the basic model (6.10). The answer depends on the magnitude of the effect, and the use to be made of derived measures like roh. With a given population, type of units, frame and method of sampling available, similar variability in sample sizes will be encountered in all designs, and the separation of its effects may not be important for 'portability'. However there are situations in which taking into account the effect of variation in cluster sizes represents a significant improvement to the basic model (6.10). One important case is the design in which greatly different cluster sizes are used in different domains. Consider the simple case in which the average cluster sizes $\bar{b}_1$ and $\bar{b}_2$ differ greatly between two domains of relative sizes W_1 and W_2 respectively. In place of the simple average of the cluster sizes, it is more appropriate to use the weighted average when combining the two domains (Kish et al, 1976):

$$\bar{b}' = \frac{W_1.\bar{b}_1 + W_2.\bar{b}_2}{W_1 + W_2}$$

In a self-weighting sample we have

$$W_1 \propto n_1 = a_1.\bar{b}_1$$

$$W_2 \propto n_2 = a_2.\bar{b}_2$$

where n_1 is the sample size and a_1 the number of primary selections in domain 1; similarly for domain 2. This gives:

$$\bar{b}'_1 = \frac{a_1.\bar{b}_1^2 + a_2.\bar{b}_2^2}{a_1.\bar{b}_1 + a_2.\bar{b}_2}$$

which is the same as (6.11) assuming a uniform cluster size within each domain.

6.5.3 INCORPORATING THE EFFECT OF WEIGHTING

As described in Section 5.3, the effect of 'haphazard' weights is to inflate variances (and defts) by a factor which tends to be similar and persist across different variables and subclasses. (Several illustrations of this persistence are given below.)

By rewriting (6.10) as equation (6.8) given earlier, we obtain a more portable roh which is not affected by the effect of haphazard weights on variance:

$$deft^2 = D_w^2[1 + (\bar{b} - 1).roh]$$

The loss factor due to weighting, D_w, is defined in equation (5.10).

6.5.4 VARIABILITY OF COMPUTED MEASURES

It is important to emphasise that variance estimates from a sample are themselves subject to variability, particularly in many practical designs based on relatively small numbers of PSUs (Section 4.4.5). Hence averaging of defts and rohs resulting from individual computations is necessary. In the present context, the important point is that roh, while in principle more portable than deft, tends to be more unstable. This is because roh is computed as the ratio of two quantities which may both be small, especially when $\bar{b}$ and hence deft is close to 1:

$$roh = \frac{deft^2 - 1}{\bar{b} - 1} \qquad (6.12)$$

Since often deft and roh values vary greatly by variable within the same survey, averaging over variables is generally not appropriate, except sometimes over a set of substantively similar variables. Stable patterns are more readily encountered when considering the same variable or similar sets of variables over diverse subclasses of the sample. Over appropriately defined sets, averaging of deft (rather than $deft^2$) values is usually preferable because of the smaller range of values encountered. An alternative sometimes useful is to fit models to a function of ($deft^2$-1), which actually is similar to modelling or averaging in terms of roh values. Several examples of this appear in the models discussed in the following sections.

Since empirically roh values tend to be rather unstable, it is often preferable to use the median rather than the mean in averaging the results, so as to limit the effect of extreme values; or one may apply an equation like (6.12) to already averaged deft values to compute roh.

ILLUSTRATION 6F EXAMPLES OF DEFTS AND ROHS FROM HOUSEHOLD SURVEYS

It is important to document information on deft and roh values for different situations, types of variables, designs, populations and subclasses. This illustration provides a number of examples, and much more needs to be done in this respect, especially from diverse surveys in developing countries.

Persistence of the Effect of Weighting

Tables 6F.(1)-(3) show that the simple expression (5.10) works well to isolate the effect of haphazard weights for diverse variables, not only for the total sample but also for cross-classes of various types and sizes cutting across the sample structure. Table 6F.(1) shows, for each variable, the ratio (D_W) of (i) the standard error for a random sample of elements with the actual sample weights, to (ii) the standard error corresponding to a self-weighting simple random sample of the same size. (See Table 4A.(3) for the variable names.) The computation is based on equation (5.12), which is distinct for each variable, but the results are mostly very close to the overall value predicted by the much simpler and general expression (5.10)

Table 6F.(2) presents two sets of results for Thailand, for which the effect of weighting, as shown in the previous table, is relatively large. The first part shows the D_W values for the urban and rural domains and for various administrative regions of the country, averaged over groups of variables. Because of the smaller sample sizes per domain, the results are generally less stable, but with the exception of the small group '6' concerning background characteristics, the results are similar across variables within domains. The second part of the table shows the same calculation for age groups, which are cross-classes, well distributed over the total sample. The effect of weighting is similar to that for the total sample, and again stable across different variables.

From a different set of surveys, Table 6F.(3) shows defts for very small cross-classes, for which the design effects are expected to be close to 1.0 except for the effect of weighting. (See section 6.5 for a discussion of subclass defts.) Values have been averaged over groups of variables individually listed in Table 4A.(1). In all cases, the results are again very close to the prediction of equation (5.10).

TABLE 6F.(1). The effect of weighting on deft values for diverse variables. Egypt and Thailand -total sample. (Source: Aliaga and Verma, 1991.)

variable			EGYPT wted deft	unwted deft	ratio	THAILAND wted deft	unwted deft	ratio
1	1	BBEFXX						
1	2	CDEAD						
1	3	CEB	1.46	1.41	1.03	1.84	1.48	1.25
1	4	CEB40	1.57	1.52	1.03	1.59	1.33	1.20
1	5	CMAR						
1	6	CSUR	1.37	1.33	1.03	1.73	1.40	1.24
1	7	EXPOS						
1	8	PREG	1.20	1.16	1.03			
1	9	SINGLE						
2	10	ATTE	2.49	2.43	1.03	2.17	1.78	1.22
2	11	BCG	1.62	1.57	1.03			
2	12	COUGH						
2	13	DIAR	1.14	1.11	1.03	1.48	1.21	1.22
2	14	DIATR	1.23	1.20	1.03	1.17	0.94	1.25
2	15	DPT	1.42	1.37	1.03			
2	16	FEVER						
2	17	FULLIM						
2	18	HCARD	0.99	0.96	1.03	1.63	1.40	1.16
2	19	HTAGE						
2	20	MEASLE	1.54	1.49	1.03			
2	21	POLIO	1.34	1.30	1.03			
2	22	TETA	1.83	1.79	1.02	1.83	1.48	1.24
2	23	TREATC						
2	24	TREATF						
2	25	WTAGE						
2	26	WTHGT						
3	27	DELAY	1.11	1.08	1.02	1.41	1.17	1.20
3	28	IDEAL	1.64	1.59	1.03	2.35	1.73	1.36
3	29	NOMORE	1.50	1.46	1.03	1.37	1.13	1.21
4	30	CUSE	2.27	2.20	1.03	1.54	1.27	1.21
4	31	EVUSE	2.70	2.63	1.03	1.72	1.41	1.22
4	32	KANY						
4	33	KMOD						
4	34	KSOURC	1.66	1.60	1.03			
4	35	UCOND	1.32	1.28	1.03			
4	36	UIUD	1.89	1.82	1.04	2.10	1.54	1.36
4	37	UMOD						
4	38	UPIL	1.66	1.62	1.03	1.86	1.55	1.20
4	39	USTER				1.88	1.56	1.21
5	40	ABST	1.20	1.17	1.03	1.25	1.08	1.16
5	41	AMEN	1.24	1.21	1.03	1.36	1.12	1.21
5	42	BF	1.24	1.20	1.03	1.51	1.26	1.20
5	43	UABST						
5	44	UTRAD						
6	45	EDUC	3.07	2.99	1.03	2.10	2.06	1.02
6	46	MBEFXX	2.24	2.18	1.03			
6	47	NOED	2.96	2.88	1.03			
				average=	1.03		average=1.22	

TABLE 6F.(2) Pervasiveness of the effect of weighting for defts over subclasses:
Illustration from Thailand.

Deft values including the effect of weighting -
averaged over variable groups, for various domains and subclasses

variable group	TOTAL	URBAN	RURAL	REG1	REG2	REG3	REG4	REG5	15-19	20-24	25-29	30-34	35-39	40-44	45-49
1	1.72	1.35	1.59	1.31	1.79	1.36	2.13	1.30	1.20	1.34	1.49	1.59	1.59	1.33	1.53
2	1.65	1.16	1.55	1.13	1.27	1.51	1.87	1.54	1.23	1.40	1.37	1.30	1.19	1.26	1.18
3	1.71	1.15	1.62	1.12	1.43	1.42	1.94	1.40	1.26	1.22	1.37	1.14	1.10	1.07	1.16
4	1.82	1.21	1.73	1.11	1.19	1.67	1.84	1.66	1.30	1.51	1.38	1.29	1.33	1.35	1.17
5	1.37	1.09	1.28	1.04	1.18	1.26	1.44	1.19	1.24	1.31	1.20	1.21	1.27	1.27	1.10
6	2.10	2.09	1.89	2.10	2.20	1.66	2.41	1.99	1.17	1.28	1.34	1.39	1.24	1.40	1.04
all	1.69	1.24	1.59	1.19	1.39	1.48	1.87	1.48	1.25	1.37	1.36	1.29	1.27	1.28	1.23

Deft values excluding the effect of weighting -

1	1.40	1.28	1.57	1.22	1.64	1.26	1.30	1.39	0.99	1.15	1.11	1.23	1.19	1.16	1.30
2	1.36	1.11	1.16	1.35	1.43	1.44	1.13	1.36	1.06	1.16	1.14	1.07	1.04	1.00	1.03
3	1.34	1.10	1.22	1.28	1.32	1.34	1.11	1.35	1.07	1.11	1.21	1.06	1.02	1.03	1.09
4	1.47	1.09	1.02	1.49	1.52	1.60	1.17	1.49	1.09	1.20	1.08	1.05	1.11	1.09	0.99
5	1.15	1.03	1.07	1.15	1.16	1.11	1.05	1.14	1.06	1.09	1.01	1.04	1.04	1.03	0.97
6	2.06	2.05	1.86	1.77	2.55	1.99	2.02	1.73	1.04	1.18	1.27	1.42	1.36	1.27	1.22
all	1.39	1.16	1.22	1.35	1.48	1.42	1.19	1.38	1.06	1.15	1.12	1.10	1.09	1.07	1.09

The effect of weighting across variables and subclasses (ratio of the above) -

1	1.23	1.05	1.01	1.07	1.09	1.08	1.64	0.94	1.21	1.16	1.34	1.29	1.34	1.15	1.18
2	1.21	1.05	1.34	0.83	0.88	1.05	1.66	1.14	1.16	1.21	1.20	1.22	1.15	1.26	1.15
3	1.27	1.05	1.33	0.87	1.09	1.06	1.75	1.03	1.18	1.10	1.13	1.07	1.08	1.04	1.06
4	1.24	1.11	1.69	0.75	0.79	1.04	1.57	1.12	1.19	1.25	1.28	1.23	1.20	1.23	1.18
5	1.19	1.06	1.20	0.91	1.02	1.14	1.37	1.04	1.17	1.20	1.19	1.16	1.21	1.24	1.13
6	1.02	1.02	1.02	1.18	0.86	0.83	1.19	1.15	1.13	1.08	1.05	0.98	0.91	1.10	0.85
All	1.21	1.06	1.30	0.88	0.93	1.05	1.57	1.07	1.18	1.19	1.21	1.18	1.17	1.20	1.12

TABLE 6F.(3). DEFTS FOR VERY SMALL CROSS-CLASSES AND DIFFERENCES BETWEEN CROSS-CLASSES FOR WEIGHTED SAMPLES.
(Source: Verma, Scott, and O'Muircheartaigh, 1980.)

Deft values for small subclasses and subclass differences, compared to estimated increase in standard error due to departure from self-weighting

Country	Indonesia									Sri Lanka									Bangladesh	
Domain	*Urban*			*Rural*			*Total*			*Urban*			*Rural*			*Total*[4]			*Rural*	*Total*
Effects of weighting[1]	*1·06*			*1·12*			*1·18*			*1·19*			*1·09*			*1·11*			*1·00*	*1·06*
Subclass difference[2]	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(1)
Subclass $\bar{b}$[3]	1·9	1·4	1·1	3·1	2·2	2·5	2·6	1·9	2·1	1·5	1·0	1·1	1·7	1·1	1·0	1·7	1·1	1·1	2·4	2·1
Deft by variable group																				
Nuptiality	1·07	1·01	1·04	1·19	1·14	1·13	1·25	1·20	1·21	1·17	1·19	1·23	1·02	1·15	1·08	1·07	1·08	1·09	0·99	1·05
Fertility	1·03	1·05	1·00	1·15	1·16	1·21	1·21	1·21	1·26	1·20	1·16	1·17	1·05	1·06	1·13	1·10	1·09	1·12	1·00	1·05
Preferences	1·02	1·02	1·06	1·21	1·19	1·25	1·25	1·23	1·29	1·19	1·15	1·26	1·07	1·07	1·10	1·13	1·10	1·12	0·99	1·04
Knowledge	1·23	1·10	0·96	1·20	1·11	1·12	1·27	1·17	1·21	1·24	1·20	1·25	1·14	1·08	1·08	1·16	1·09	1·16	1·04	1·09
Use	0·99	1·02	1·08	1·11	1·14	1·07	1·15	1·19	1·13	1·16	1·14	1·24	1·02	1·06	1·08	1·06	1·08	1·12	1·04	1·04
Average *deft* (all variables)	1·05	1·04	1·03	1·16	1·15	1·16	1·21	1·20	1·21	1·19	1·16	1·23	1·05	1·06	1·09	1·09	1·09	1·12	1·01	1·05

[1] Increase in standard error due to departure from self-weighting within country or domain

[2] (1): Subclass "age 45-49".
(2): Difference between subclasses "age 35–44" and "age 45-49"
(3): Difference between subclasses "marriage duration 0-4 years" and "marriage duration 5-9 years"

[3] For differences, $\bar{b}$ is defined as half the harmonic mean of $\bar{b}_i$ for the two subclasses.

[4] Includes the small "Estate" domain.

Dependence on Rohs on the Nature of the Variable.

It is important to note that deft and roh value are specific to the substantive variable under consideration. Different variables in the same survey may have several-fold differences in roh values. Hence there is not one 'design effect' for a sample, but different factors depending on the variable.

In Table 6F.(4), Kalton and Blunden (1973) report markedly different design effects for individual and household characteristics in the British General Household survey. The implied roh values (not shown here) differ even more markedly.

As another example, Kish et al (1976) report greatly differing roh values for different types of variables in fertility surveys. It is found in fertility surveys that roh and deft values are often low for demographic variables (reflecting a low degree of homogeneity among individuals within clusters), higher for socio-economic variables and still higher for variables pertaining to household characteristics, and particularly housing conditions.

Table 6F.(5) presents the results from two more systematic comparative studies on the same topic, covering a common set of variables from similar fertility surveys in a number of developing countries. The values shown are averaged over sets of variables in the same substantive group.

The top panel shows some results from the World Fertility Survey. In this panel, median rather than mean values for rohs over groups of variables are shown, since the medians are less affected by extreme values in the computed results.

The middle panel of Table 6F.(5) shows a similar set of results from the Demographic and Health Surveys covering a number of developing countries. (See Tables 4A.(3) and (4) for names of the countries and variables included in this table.) The results for rohs, averaged over countries for groups of variables, are remarkably similar for the two sets, despite the differences between them in terms of the countries covered, sample designs, and to some extent, the actual variables covered. This is summarised in the bottom panel of the table.

TABLE 6F.(4). A comparison of defts for different types of variables.

Comparison of $\sqrt{\text{deff}}$'s for a selection of characteristics for Great Britain, the Greater London Council area and Scotland

Characteristic	$\sqrt{\text{deff}}$		
	Great Britain	G. L. C.	Scotland
Head of household:			
is chronic sick	1.31	1.37	1.24
is working	1.29	1.27	1.19
Single-person household	1.15	1.28	0.78
Household has no car	1.57	1.53	1.77
Household's accommodation:			
Pre-1919 building	2.25	2.22	1.38
Owner-occupied	2.01	2.03	1.99
Has a fixed bath	1.92	1.44	1.18
Has an inside lavatory	2.07	1.90	1.40
Has fixed central heating	1.72	1.67	1.07
Is a whole detached house	1.91	2.69	1.55

TABLE 6F.(5). ROH VALUES FOR GROUPS OF VARIABLES: COMPARISON ACROSS COUNTRIES FROM THE WFS AND DHS SURVEY PROGRAMMES.
(Sources: Verma et al, 1980; Aliaga and Verma, 1991.)

[1] Median roh values from some WFS surveys

Median roh by variable group and country

	Mexico	*Peru*	*Jamaica*	*Indonesia*	*Sri Lanka*	*Thailand*	*Guyana*	*Nepal*	*Colombia*	*Bangladesh*	*Fiji*	*Costa Rica*	*All†*
Nuptiality	0·03	0·01	0·04	0·02	0·02	0·01	0·02	0·02	0·01	0·01	0·02	0·01	0·02
Fertility	0·06	0·04	0·03	0·02	0·03	0·01	0·05	0·02	0·00	0·01	0·00	0·02	0·025
Preferences	0·03	0·05	0·07	0·05	0·03	0·02	0·03	0·05	0·01	0·02	0·02	0·00	0·03
Knowledge	0·17	0·14	0·09	0·14	0·07	0·08	0·08	0·05	0·05	0·06	0·04	0·01	0·08
Use	0·11	0·09	0·04	0·06	0·07	0·06	0·03	0·03	0·04	0·03	0·02	0·02	0·05
All variables†	0·07	0·05	0·05	0·05	0·05	0·04	0·03	0·03	0·02	0·02	0·02	0·01	0·04

† Average of median values. Countries arranged according to last row.

TABLE 6F.(5) continued.

[2] Roh values from some DHS surveys

variable group	COUNTRY: 1	2	3	4	5	6	7	8	9	10	11	12
1	.011	.003	.009	.007	.019	.021	.025	.022	.025	.033	.021	.045
2	.026	.017	.014	.022	.024	.021		.030	.072	.023	.008	.038
3	.051	.021	.005	.012	.008	.037	.033	.024	.007	.036	.036	.043
4	.082	.019	.043	.045	.017	.067	.023	.065	.074	.071	.097	.055
5	.025	.007	.008	.008	-.009	.013	.011	.011	.013	.015	.003	.012
6	.050	.062	.064	.029	.132	.080	.225	.163	.080	.138	.098	.088
all	.034	.016	.019	.019	.021	.032	.029	.043	.038	.041	.032	.042
b-bar	29.9	25.2	24.1	28.2	12.8	12.8	17.2	39.3	24.5	23.5	16.1	23.5

[3] Comparison of averaged roh values; selected WFS and DHS surveys.

Variable group	deft		roh	
	WFS	DHS	WFS	DHS
non-contraceptive factors				
affecting fertility		1.20		.010
nuptiality	1.30	1.25	.020	.020
fertility	1.25	1.25	.025	.020
child health		1.35		.015
fertility preferences	1.40	1.35	.030	.025
contraceptive use	1.55	1.60	.050	.050
contraceptive knowledge	2.00		.080	
background characteristics		1.85		.095

The Effect of Cluster Sizes

Table 6F.(6) shows the effect of size and nature of the PSUs on roh values. In practical design work, an important issue concerns the choice of the number of stages in which the sample should be selected. In many situations, the frame defines some 'basic' area units which can be used for sample selection and organisation of fieldwork, and the issue is whether such units should be selected directly as the PSUs or clustered further by the introduction of higher stages (International Statistical Institute, 1975). The type of units which form the PSUs will vary greatly depending on how this decision is taken. In this example, the effect of the type of units is considered by computing sampling errors by taking three entirely different types of units as the PSUs in the same survey, using the methodology described in Section 5.5: (i) very large units with an average size of 100,000 households; (ii) large units of average size 5,000 households; and (iii) compact units of average size around 250 households each. This gave three designs:

[1] A four stage design with (i) as PSUs, (ii) as SSUs, (iii) as the ultimate area units, followed by the sampling of households;
[2] a three stage design with (ii) as PSUs and (iii) as SSUs; and
[3] a two stage design with (iii) as PSUs.

The $\bar{b}$ values for the three designs were around 84, 42 and 14 respectively. The table shows the computed deft and roh averaged over a number of variables. Clearly roh declines as the size of the PSUs increases. Actually the real intraclass correlations differ much more markedly between the three types of units than reflected by the figures in the table. This is because of heavy clustering of the subsamples within PSUs in case (i), and also to some extent in case (ii).

TABLE 6F.(6). Defts and rohs in designs with different types and sizes of units

Design	[1]	[2]	[3]
b-bar	84	42	14
deft	1.66	1.55	1.39
roh	0.021	0.034	0.072

The Method of Subsampling within PSUs.

Kalton (1979) gives the following example illustrating the impact of the method of subsampling with PSUs.

> "Two British surveys employed similar sample designs at the early stages, both taking stratified PPS samples of local authority areas at the first stage with similar stratification factors, and PPS selections of two wards per selected area at the second stage. In the first survey a systematic sample of electors was then drawn throughout each selected ward while in the second survey a further stage of clustering was employed, taking 10 electors by means of a 1 in 12 systematic sample from a random start with each ward. Both surveys collected data on respondents' terminal ages of education. For the first survey, for the proportion of respondents finishing their education at 16 or under, the estimated average intraclass correlation within strata was 0.04; for the second survey, for the proportion finishing their education at 15, it was 0.26."

The explanation for the very much higher degree of homogeneity in the second survey lies mainly in the compactness of subsampling within PSUs.

6.6 MODELLING SAMPLING ERRORS FOR SUBCLASSES

6.6.1 REQUIREMENTS

This section describes how, for a given substantive variable or group of similar variables, standard errors over diverse subclasses of interest may be related to the more readily computed errors with the total sample as the base. Establishing such relationships is one of the primary objectives of modelling sampling errors in large-scale multipurpose surveys. The measures deft and roh provide empirically useful means of achieving this.

Several important points should be noted in relation to subclass sampling errors.

(1) National surveys are frequently required to provide separate results for different geographical domains of the country, which may differ in population characteristics and sample design. For each domain the results are usually required in full detail for numerous cross-classifications of the population elements. Full-scale computation of sampling errors is difficult for every domain because of the volume of the work involved. Also the results for individual domains are subject to greater instability because of the smaller number of PSUs selected in individual domain samples. It is therefore most useful to be able to express domain sampling errors (se_g) in terms of total sample errors (se_t).

(2) In addition, many subclasses of interest are formed by crossclassifications in terms of individual characteristics of households or persons (Section 4.3.2). Such subclasses may be more or less uniformly distributed over the population. In so far as any subclasses are evenly spread over sample clusters, the effective cluster sizes go down roughly in proportion to subclass size. This tends to attenuate the effects of clustering and stratification on the magnitude of the sampling error, thereby requiring a different type of modelling than that for geographic domains (1).

(3) In practice of course, subclasses are rarely uniformly distributed. As a result coefficients of variation of effective cluster sizes tend to be higher for subclasses than for the total sample where these variations are more readily controlled by the design. The resulting increase in error variance needs to be taken into account. Incidently, this also tends to increase the bias in ratio estimation for subclasses.

(4) The modelling of subclass and domain errors in terms of total sample errors should involve parameters which are readily available or estimated. Simple models have the advantage of practicality, but sometimes they are inadequate for providing the required degree of accuracy. In practice a compromise is required between the requirements of simplicity and accuracy. One should seek the simplest model which meets the minimum acceptable standards of accuracy; any complexity should be entertained only if it is clearly justified and unavoidable in the light of the objectives for which the information on sampling errors is produced.

(5) In addition to subclasses, sampling errors are also required for subclass differences and other more complex measures of relationship. This modelling is related to that for subclass sampling errors will be considered in Section 6.7.

6.6.2 THE BASIC MODEL

In discussing the relationship, for a given variable or a set of similar variables, of the sampling error for geographical domains or other subclasses to the error estimated with the total sample as the base, the basic model is obtained by putting together (6.6) and (6.8):

$$se^2 = \left(\frac{s^2}{n}.D_W^2\right).[1 + d^2] \tag{6.13}$$

where

$$d^2 = \left(\frac{deft}{D_W}\right)^2 - 1 , \tag{6.14}$$

and n is the sample size, s^2 the population variance and D_W^2 the effect of weights on the variance, for the total sample or some other base such as a particular domain or subclass.

The first factor in (6.13) is the SRS variance (s^2/n), modified by the effect of weighting as described in Section 5.3. The second factor (in square brackets) is the design effect, apart from the effect of weighting. Of course, d^2 can also be written in terms of the average cluster sizes (with or without the modification (6.11)) and roh values as

$$d^2 = (\bar{b}' - 1).roh \tag{6.15}$$

Similar expressions apply for the total sample and any particular domain or subclasses. The relationship between their variances is determined by how the various components (namely n, s^2, D_W^2, $deft^2$, cluster size, and roh) relate as we move from the total sample to particular domains or subclasses. In the following subsections, some specific situations are considered in detail.

6.6.3 SAMPLING ERRORS FOR GEOGRAPHICAL DOMAINS

The objective here is to sketch out some specific models which have been found useful in practice. Several situations may be distinguished.

(1) If the nature of the population and the sample design are similar across different domains, one may begin with the assumption that deft and population variance (s^2) are also similar, thus making the standard error of a given statistic inversely proportional to the square-root of the corresponding sample size:

$$se_g = se_t \sqrt{\frac{n_t}{n_g}} \tag{6.16}$$

where subscript g refers to a particular geographical domain and t to the total sample.

(2) When the size of the estimate $\bar{y}$ differs significantly by domain, it is often more reasonable to assume that the coefficient of variation $s/\bar{y}$ rather than standard deviation (s) is constant across domains. This implies that (6.13) applies in terms of relative errors

$$rse_g = rse_t \sqrt{\frac{n_t}{n_g}} \tag{6.16'}$$

It is possible to incorporate more complicated forms of variation of s (or cv) with $\bar{y}$, but in many situations it is not worthwhile to do so. For a proportion (p) in any case, it is possible to incorporate the ratio of domain to total population standard deviation by introducing on the right hand side of (6.13) the factor

$$\left[\frac{p_g.(1-p_g)}{p_t.(1-p_t)}\right]^{1/2}$$

Again this factor tends to be close to 1.0 except for large differences in p - which do not usually occur across geographical domains of the same population.

(3) Usually it is much more important to incorporate into (6.13) the effect of differences in domain and total sample defts. Referring to model (6.8), these may arise due to (i) differences in the effect of weighting; (ii) differences in the cluster sizes; and (iii) differences in other aspects of the design which affect roh. All these differences can be important. For instance, weights due to disproportionate allocation are often introduced across rather than within domains (eg oversampling of urban areas but self-weighting samples within urban and rural areas separately); this would make all $D_g = 1$ but $D_t > 1$. (D_g and D_t refer to the loss due to weighting (D_W) for a domain, g, and for the total sample, t, respectively.) Similarly there are often good reasons to choose markedly differing designs and sample takes in urban and rural areas. The relation between the design effects follows from (6.13) as:

$$\frac{[(deft_g/D_g)^2-1]/(b_g'-1)}{[(deft_t/D_t)^2-1]/(b_t'-1)} = \frac{roh_g}{roh_t} = c_g, \quad say \tag{6.17}$$

The $\bar{b}'$ may be simply the mean cluster size, or the modified version (6.11) if those sizes vary significantly within and across domains. Factors like $\bar{b}_g'$ and D_g are easily obtained for all (usually a small number of) domains as they are roughly independent of the variables concerned. The roh values depend on the variable concerned, but the ratio (roh_g/roh_t) can be expected to be similar for different variables, all subject to the same differences in domain designs. This suggests the following simple model. The model is to compute the left hand side of (6.17) for each variable in a set, and take its average (say c_g; mean or median) as an estimate of the overall value of the ratio (roh_g/roh_t) applicable to all variables in the set. Using this common c_g, (6.15) can be applied in reverse to obtain the modelled value of $deft_g$ from $deft_t$ for any particular variable. The relationship may be expressed simply as a set of predicted $deft_g$ versus $deft_t$ graphs, one for each domain. Finally the 'predicted' ratio ($deft_g/deft_t$) can be introduced into the right hand side of (6.16) or (6.16') to incorporate into the model the effect of differences in domain and total sample defts.

A cruder but more robust alternative to (6.17) is to seek a relationship in terms of the averaged ratio of defts, in place of the generally less stable roh values:

$$deft_g/deft_t \;=\; c_g\,, \tag{6.17'}$$

where c_g is a constant for a given domain, an averaged value of the ratio of the domain to total sample defts for diverse variables.

6.6.4 SAMPLING ERRORS FOR DISTRIBUTED SUBCLASSES

There are two major opposing factors to be considered as we move from the total sample to a subclass distributed across the population: (i) Variance is increased in proportion to the reduction in the effective sample size. (ii) However, variance is reduced since the design effect also goes down as a result of the reduction in the effective cluster sizes. In addition, (iii) subclasses may also differ from the total population in terms of their population variances (or cv's). Factors (i) and (iii) affect the SRS variance and are discussed first briefly. Factor (ii) concerns the total sample and subclass defts, and requires more detailed consideration.

Sample Size

The effect of the sample size is large and straightforward: variance increasing as ($1/m_s$), where m_s is the size of the subclass as a proportion of the total sample. The model can be made more elaborate by incorporating the effect of differences between subclasses in the population variances and sizes of the estimates involved, if such differences are found to be sufficiently important.

Population Variance (or Coefficient of Variation, cv)

As in the case of many geographical domains, it is often reasonable to assume that for a given variable or characteristic, the coefficient of variation remains fairly stable across distributed subclasses as well. Generally however, variations in cv's are more likely to be important in the case of subclasses. This is because characteristics defining distributed classes are often closely related to the substantive variables of interest. For example, the patterns of household consumption (and their variability) are related to household size and demographic composition which form common classifications (subclasses) in the tabulation and analysis of survey data. Similarly, labour force, fertility and many other variables are closely related to age-sex classifications. As a result, within any classification category (subclass), variables related to the classifier tend to have lower cv values.

<u>Subclass Design Effects</u>

The major impact of moving from the total sample to subclasses is to reduce the effective cluster size, thus reducing deft in accordance with equation (5.15). For subclasses distributed over most sample areas, the effective cluster size declines in proportion to size of the subclass:

$$\bar{b}_s = \bar{b}_t . m_s ; \quad m_s = \frac{n_s}{n_t}$$

where m_s is the size of the subclass as a proportion of the total sample, subscript s refers to quantities for a subclass and t to the corresponding quantities for the total sample. For very small subclasses, deft values may approach 1.0 except for the effect of random weights (Section 5.3) which tends to persist across variables and subclasses independently of other aspects of the design. By introducing the notation

$$d^2 = (deft/D)^2 - 1; \quad m_s = n_s/n_t; \quad b_s = b_t . m_s \tag{6.18}$$

where 'D' refers to the effect of weighting on deft, the basic model can be written as

$$\frac{d_s^2}{d_t^2} = \frac{b_s - 1}{b_t - 1} . \frac{roh_s}{roh_t} \tag{6.19}$$

The main assumption to consider is the relationship between subclass and total sample roh values.

Often, the type of units and methods of subsampling are essentially the same whether we consider the total sample or reasonably distributed subclasses - making their roh values similar for a given variable. The simplest model is to assume roh_s and roh_t to be the same; with the added assumption that the subclass and the cluster sizes are not too small, we obtain for cross-classes:

$$\frac{d_s^2}{d_t^2} = \frac{b_s}{b_t} = m_s \tag{6.20}$$

It has been found that the decline in deft, with decreasing subclass size (m_s) tends to be less rapid in practice than that implied by the above. The main source of this effect is that subclasses are hardly ever uniformly distributed across sample clusters, making cluster sizes more variable for subclasses than they are for the whole sample. This results in increased subclass deft, above that given by (6.20), which can be modelled in several ways.

Model [1]

One model to account for the higher subclass deft than implied by (6.20) is to assume that $roh_s > roh_t$. A simple approximation has been to assume that the effect can be taken into account reasonably well by taking roh_s to be higher than roh_t for a given variable by some common factor k greater than 1.0, giving

$$\frac{d_s^2}{d_t^2} = k.\frac{b_s-1}{b_t-1} \tag{6.21}$$

Factor k may be determined empirically by fitting the model to computed results for subclasses. In principle it is possible to determine different values of k for different groups of variable and/or subclasses. From a large set of computations Kish et al (1976) have suggested a value k = 1.2 for general use.

Model [2]

Alternatively, or in addition, one may also incorporate the effect of increased variability in cluster sizes as we move from the total sample to subclasses, ie replace the simple average of cluster sizes by the modified quantity

$$\bar{b}' = \bar{b}.[1 + cv(b_j)^2] \tag{6.22}$$

reflecting the variability in cluster sizes (Section 6.5.2). The effect of this modification is similar to assuming k>1 as in the previous case.

Model [3]

Another proposal has been to argue that (6.20) as an approximation to (6.19) represents one extreme (namely that of a pure cross-class), while the other extreme is represented by completely segregated classes (like geographical domains) for which it is reasonable to assume that $deft_s$ is similar to $deft_t$, giving

$$\frac{d_s^2}{d_t^2} = 1 = m_s^o$$

Hence, the general form for classes which are neither completely separated nor distributed entirely uniformly, may be written as

$$\frac{d_s^2}{d_t^2} = m_s^{\alpha}; \quad 0 \le \alpha \le 1 \tag{6.23}$$

with the parameter α to be empirically determined for different types of subclasses, expected to be in the range (0-1), with values at the upper end corresponding to cross-classes and at the lower end to highly segregated or geographical classes.

The last mentioned model has been developed and tested on the basis of sampling errors computed for different types of variables over many subclasses in a number of surveys by Verma et al (1980). It was found that α values were generally larger (around 0.7 to 1.0) for well-distributed demographic classes, but smaller (around 0.4-0.7) for the less well-distributed socioeconomic classes. In a similar study (Aliaga and Verma, 1991), the overall value of α for age groups (which are practically true cross-classes) was found to be close to 1 (=0.97), with the subclass results closely predicted (R^2 = 0.78) by (6.23).

Model [3] has the advantage over model [1] in that it satisfies the boundary condition $deft_s = deft_t$ when m_s approaches 1 (total sample) irrespective of the value of parameter α; and over model [2] in that it includes an empirically determined parameter which can take account of the results of actual computation for different types of subclasses. In common with other models, it of course assumes that for a given degree of 'cross-classedness' the substantive nature of the variable as well as of the characteristic defining the subclass need not be considered. This however is most unlikely to be the case, and thus it would be more appropriate to estimate α separately for similar groups of variables within subclass groups of similar size and cross-classedness within each survey.

Relationship in terms of standard errors

The various ideas discussed in this section can be brought together to model actual variances or standard errors for domains and subclasses in terms of the same for the total sample. This is done by modelling the various components of the basic expression (6.13) on the basis of appropriate assumptions. For example, assuming that in moving from the total sample to a subclass, the population variance s^2 and the effect of weighting D_W^2 remain essentially unchanged, and using model [3] to relate the design effects, we can express the total sample (t) and subclass (s) variances as follows.

$$se_t^2 = \left(\frac{s^2}{n_t}.D_W^2\right).\left[1 + d_t^2\right]$$

$$se_s^2 = \left(\frac{s^2}{n_s}.D_W^2\right).\left[1 + m_s^\alpha.d_t^2\right] = se_t^2.\left(\frac{1}{m_s}\right).\left[\frac{1+m_s^\alpha.d_t^2}{1 + d_t^2}\right] \tag{6.24}$$

In the expression on the extreme right, the factor $(1/m_s)$ is the increase in SRS variance due to the reduced sample size from n_t to $n_s=m_s.n_t$, and the last factor is the compensatory effect of reduced design effect.

For the special case of a true cross-class (α=1), the above becomes

$$se_s^2 = \left(\frac{s^2}{n_s}.D_W^2\right).\left[1 + m_s.d_t^2\right] = se_t^2.\left(\frac{1}{m_s}\right).\left[\frac{1+m_s.d_t^2}{1 + d_t^2}\right] \tag{6.25}$$

ILLUSTRATION 6G PATTERNS OF VARIATION OF SUBCLASS DEFTS

Defts for geographical domains

As an application of model (6.17), Table 6G.(1) has been constructed from the results of a survey in Turkey (Turkey 1980; the results here are taken from Verma 1982). The sample was selected from eight geographical domains defined in terms of locality size varying from large metropolitan areas (domain 1) to small villages (domain 8). In domains (1)-(5), the sample was selected in three stages: wards with PPS as PSUs; two blocks per ward as SSUs; and finally a self-weighting sample of households. The computed and averaged c_g values varied from 0.4 to 1.5 across the domains corresponding to a range of around 1.20 to 1.65 for the overall averaged deft_g values.

The generally increasing values of c_g from domain (1) to domain (5) reflect the effect of increasing compactness of the sampling units (decreasing ward size with decreasing size of the locality). In rural domains (6) - (8), the sample was selected in two stages: selection of villages, following by sampling of households within selected villages. Villages were selected with constant probability - rather than with PPS - within each domain (village size group) because no reliable measures of size were available. The higher c_g values in particular for domains (6) and (8) reflect the variability in cluster sizes as a result of this selection procedure. For reference, deft_g values averaged over 27 substantive variables for which sampling errors were computed are also shown in the table. The table also provides information of goodness of fit of the model in this particular example, and compares some predicted and actually computed values. In applying the model, it was assumed that population variance varied as equation (5.5) for proportions and proportionally to $\bar{y}$ for means (Section 6.4). The agreement between the predicted and computed standard errors by domain is generally good.

Table 6G.(2) provides an example of defts from the total sample, geographical domains and some distributed subclasses from a number of fertility surveys, reference to which has already been made in other illustrations. The deft values have been averaged over groups of similar variables. In several of the countries shown, lower defts found in the urban sector reflect the use of smaller cluster sizes in comparison with rural areas. The same effect is seen more markedly when we consider cross-classes such as age groups. For small subclasses deft values tend to 1.0 as implied by model (6.23). However, some differences remain by type of variable. More importantly, for non-self weighting samples (such as Indonesia, Bangladesh, Sri Lanka), the effect of weighting persists even for very small subclasses, as implied in model (6.8).

Table 6G.(3) provides an illustration from one of a series of in-depth fertility surveys conducted in various provinces and municipalities of China during 1984-86 (China 1986). In the survey shown from the Shaanxi Province, the sample was heavily clustered with five sampling stages, with 29 sample counties as the PSUs, and around 4,500 women from 6,000 households as the ultimate units. The sample was self-weighting. A model of the form (6.23) was fitted, with $\alpha = 0.6$ as the empirically determined parameter estimated from actual computations over a wide range of subclasses. The top panel of the table shows the variation of deft by subclass size for a number of variables. Combining this model with the assumption that population variance is the same in different subclasses for a given variable (eq. 6.24), the lower panel shows the predicted variation of standard error by subclass size for different variables.

TABLE 6G.(1). Variation of averaged defts across geographical domains.

Comparison of (a) computed and (b) predicted standard errors for geographic domains in the Turkish Fertility Survey

	Metro-politan	Large cities	Medium cities	Small cities	Towns	Large villages	Medium villages	Small villages	Total
Domain									
Domain size (n_g)	648	697	350	318	628	497	734	559	4431
Cluster size, $\bar{b}_g$	19.1	19.4	21.9	22.7	20.3	19.1	21.6	24.3	20.7
No. of clusters a_g	34	36	16	14	31	26	34	23	214
Average deft$_g^2$ [a]	1.21	1.32	1.57	1.33	1.50	1.65	1.47	1.66	1.48
Estimated parameter c_g	0.39	0.64	1.25	0.64	1.05	1.47	0.98	1.48	–
Goodness of fit, R_g^2	0.11	0.26	0.37	0.28	0.19	0.56	0.22	0.29	–
Variable									
Age at first marriage									
(a)	0.190	0.170	0.247	0.238	0.221	0.211	0.148	0.218	0.072
(b)	0.157	0.171	0.283	0.242	0.190	0.233	0.170	0.219	–
Children ever born									
(a)	0.111	0.111	0.215	0.241	0.178	0.158	0.102	0.247	0.060
(b)	0.114	0.121	0.207	0.196	0.165	0.213	0.162	0.204	–
Proportion who know of pill									
(a)	0.008	0.011	0.031	0.029	0.026	0.029	0.030	0.040	0.010
(b)	0.011	0.016	0.029	0.028	0.026	0.036	0.028	0.039	–
Proportion currently using contraception									
(a)	0.018	0.020	0.038	0.024	0.032	0.028	0.016	0.014	0.008
(b)	0.020	0.020	0.032	0.026	0.022	0.023	0.015	0.014	–

Source: Haceteppe Institute of Population Studies (1980)

TABLE 6G.(2). EXAMPLES OF VARIATION OF DEFTS ACROSS GEOGRAPHICAL DOMAINS AND CROSS-CLASSES FOR GROUPS OF VARIABLES.
(Source: Verma et al, 1980.)

	Nepal	Mexico	Thailand	Indonesia	Colombia	Peru	Bangladesh	Fiji	Sri Lanka	Guyana	Jamaica	Costa Rica
TOTAL SAMPLE $\bar{b}$	148.5	34.4	54.6	24.3	8.2	13.8	27.1	49.3	10.5	18.4	6.7	10.5
$\bar{b}$	163.9	34.4	77.0	25.8	36.1	14.2	28.5	49.4	10.6	18.6	6.7	13.6
Nuptiality	1.91	1.39	1.28	1.45	1.31	1.10	1.22	1.28	1.13	1.16	1.12	1.16
Fertility	1.92	1.58	1.38	1.41	1.22	1.17	1.12	1.04	1.19	1.02	1.06	1.05
Preferences	2.76	1.52	1.37	1.55	1.20	1.24	1.21	1.37	1.16	1.13	1.08	1.00
Knowledge	3.21	2.81	2.48	2.44	2.50	1.86	1.66	1.66	1.36	1.52	1.25	1.00
Use	2.23	1.92	2.15	1.70	1.70	1.36	1.31	1.31	1.30	1.24	1.07	1.07
All variables	2.32	1.71	1.61	1.62	1.49	1.27	1.26	1.26	1.22	1.17	1.10	1.06
URBAN DOMAIN · $\bar{b}$		33.2	17.5	17.6	3.5	12.2	18.6	51.9	9.0	21.1	8.2	7.3
Nuptiality		1.33	1.17	1.28	1.10	1.08	1.26	1.25	1.17	1.03	1.20	1.02
Fertility		1.41	1.45	1.22	1.02	1.04	1.09	1.15	1.17	1.00	1.08	1.07
Preferences		1.22	1.43	1.23	1.04	1.11	1.01	1.48	1.26	0.98	1.04	1.01
Knowledge		2.04	1.23	1.79	1.12	1.40	1.45	1.23	1.53	1.32	1.42	0.99
Use	–	1.56	1.42	1.43	1.12	1.18	1.24	0.96	1.33	1.35	1.09	0.99
RURAL DOMAIN . $\bar{b}$	–	36.2	87.6	29.6	57.2	17.8	31.4	47.9	11.2	17.2	5.7	20.3
Nuptiality	–	1.38	1.26	1.36	1.40	1.03	1.15	1.20	1.17	1.17	1.04	1.25
Fertility	–	1.53	1.34	1.34	1.33	1.14	1.06	1.20	1.15	1.02	1.05	1.04
Preferences		1.62	1.35	1.48	1.26	1.38	1.14	1.25	1.13	1.19	1.12	1.00
Knowledge		2.53	2.55	2.33	3.05	1.91	1.56	1.78	1.29	1.58	1.12	1.00
Use		1.88	2.29	1.63	2.24	1.63	1.29	1.43	1.29	1.14	1.03	1.15
SUBCLASS "AGE 25-34" . $\bar{b}$	50.0	12.4	19.4	8.0	2.9	4.8	8.0	20.1	4.2	6.4	2.2	4.2
Nuptiality	1.46	1.08	1.25	1.26	1.05	1.12	1.10	1.19	1.12	1.10	1.10	1.05
Fertility	1.41	1.30	1.24	1.30	1.06	1.11	1.06	1.08	1.14	1.04	1.05	1.02
Preferences	1.94	1.08	1.30	1.31	1.08	1.13	1.09	1.18	1.15	1.07	1.00	1.05
Knowledge	2.26	1.86	1.57	1.76	1.64	1.34	1.38	1.40	1.22	1.17	1.06	1.02
Use	1.76	1.54	1.51	1.43	1.37	1.15	1.17	1.16	1.24	1.07	0.94	1.02

TABLE 6G.(3). Standard errors and defts as functions of subclass size.
(Source: China, 1986.)

Variation of deft by subclass size

	sample size of the subclass:										
	3001-3500	2501-3000	2001-2500	1501-2000	1001-1500	701-1000	501-700	301-500	201-300	101-200	<100
Variable:											
reported best age at marriage	3.67	3.60	3.42	3.22	2.98	2.67	2.44	2.45	1.98	1.82	1.59
lived with parents after marriage*	2.98	2.90	2.74	2.59	2.41	2.18	2.01	1.86	1.68	1.56	1.40
number of children desired	2.32	2.24	2.14	2.04	1.91	1.75	1.63	1.54	1.42	1.34	1.23
mean age at first marriage	2.23	2.15	2.06	1.96	1.84	1.69	1.59	1.50	1.38	1.31	1.21
used contra before last pregnancy*	1.75	1.64	1.57	1.64	1.49	1.40	1.33	1.28	1.21	1.17	1.12
mean births in past 5 years	1.57	1.52	1.47	1.42	1.36	1.30	1.25	1.20	1.15	1.14	1.09
mean no. of children ever-born	1.37	1.34	1.32	1.27	1.23	1.19	1.16	1.13	1.09	1.08	1.05
births in first 5 yrs of marriage	1.28	1.26	1.24	1.21	1.17	1.14	1.12	1.09	1.07	1.05	1.04
p wanting to have more children*	1.19	1.18	1.15	1.14	1.12	1.09	1.08	1.06	1.04	1.03	1.02

Variation of standard error by subclass size

reported best age at marriage	0.12	0.13	0.13	0.14	0.15	0.16	0.18	0.20	0.22	0.25	0.31
lived with parents after marriage*	2.04	2.12	2.23	2.34	2.51	2.79	3.07	3.37	3.92	4.46	5.62
number of children desired	0.04	0.04	0.05	0.05	0.05	0.06	0.07	0.07	0.09	0.10	0.13
mean age at first marriage	0.10	0.11	0.11	0.12	0.13	0.15	0.17	0.18	0.22	0.25	0.33
used contra before last pregnancy*	1.47	1.53	1.61	1.74	1.90	2.18	2.49	2.81	3.44	4.07	5.49
mean births in past 5 years	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.06	0.09
mean no. of children ever-born	0.04	0.04	0.05	0.05	0.06	0.07	0.08	0.09	0.11	0.14	0.18
births in first 5 yrs of marriage	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.06	0.07	0.10
wanting to have more children*	0.86	0.94	1.01	1.11	1.26	1.31	1.78	2.08	2.64	3.44	4.48

* proportion

6.7 SAMPLING ERRORS FOR SUBCLASS DIFFERENCES AND OTHER COMPLEX STATISTICS

Subclass Differences

The basic model for the variance of the difference of two subclass means, v(r-r'), may be expressed as follows:

$$[v_o(r)+v_o(r')] \le v(r-r') \le [v(r)+v(r')] \tag{6.26'}$$

The first expression on the left is the variance of the difference assuming independent simple random sampling - ie disregarding design effect, but also any covariance between the classes. The last expression on the right is the sum of the subclass variances, assuming full subclass defts but no covariance. The actual variance for the subclass difference should be somewhere between these two limits. The reduction in variance due to the presence of correlation between subclasses which come from the same sample of primary units may be summarised by introducing the factor (1-R), where R is a synthetic coefficient of covariance in the comparison of subclasses. By virtue of the definition of the various terms, the following set of equalities may be written:

$$\begin{aligned} & deft_d^2.[v_o(r)+v_o(r')] \\ &= v(r-r') \\ &= (1-R).[v(r)+v(r')] \\ &= (1-R).deft_s^2.[v_o(r)+v_o(r')] \end{aligned} \tag{6.26}$$

where $deft_s$ is the common or averaged value of the design effect for the subclasses being compared. The value of R can be estimated from actual computation of var(r), var(r') and var(r-r') for a set of similar subclasses, and then used as a basis to relate these quantities for other subclasses not covered in the initial computations.

Some empirical results on the values of R encountered in household surveys are noted below. The data in Table 6H.(1) are from the British General Household Survey (Kalton and Blunden 1973). Note that the correlations are much larger for accommodation characteristics than for household characteristics. The former are also the characteristics with large design effects. This is expected because both defts and R are determined by the common effect of clustering for subclasses coming from the same PSUs. Kish (1968) reports an average value R = 0.16 from a series of surveys concerning consumer attitudes. Verma et al (1980) report values which showed systematic differences according to the nature of the subclasses being compared. For well-distributed demographic classes (such as age and sex groups), R values for diverse variables and samples tended to be in the range 0.15-0.30; while for the less well distributed socioeconomic classes (groups by occupation, level of education, race or ethnicity, etc), the values were generally much lower, mostly in the range 0.05-0.15.

TABLE 6H.(1). AN EXAMPLE OF SAMPLING ERRORS FOR DIFFERENCES.

Subclass $\sqrt{\text{deff}}$'s for a selection of characteristics: (a) for households in which the head is aged under 45 and (b) for households in which the head is aged 45 or over

Characteristic	Per cent (a) r	Per cent (b) r'	$\sqrt{\text{deff}}$ (1) All	$\sqrt{\text{deff}}$ (2) r	$\sqrt{\text{deff}}$ (3) r'	$\sqrt{\text{deff}}$ (4) $r-r'$	Corr (r, r') R
Household characteristics							
Head of household:							
is chronic sick	8.5	33.4	1.31	1.09	1.19	1.05	0.19
is a hospital out-patient	11.4	11.3	1.07	0.96	1.19	1.09	− 0.08
has made a doctor consultation	10.7	14.5	1.03	1.16	0.93	1.05	0.03
is working	93.7	56.4	1.29	1.06	1.16	1.04	0.19
Single-person household	5 6	24.9	1.15	1.15	1.12	1 10	0.05
Household has no car	33.2	53.7	1.57	1.34	1.34	1.07	0.37
Accommodation characteristics							
Pre-1919 building	24.8	34.0	2.25	1.64	1.90	1.11	0.60
Owner occupied	53.5	47.8	2.01	1.65	1.69	1.23	0.46
Has a fixed bath	95.2	90.1	1.92	1.69	1 62	1.20	0.47
Has an inside lavatory	90.0	86.1	2.07	1.61	1.77	1.19	0.50
Has fixed central heating	34.5	24.4	1.72	1.42	1.32	0.98	0.55
Is a whole detached house	14.9	16.8	1.91	1.62	1.61	1.24	0.42
Sample size			3346	1266	2077		

Subclass $\sqrt{\text{deff}}$'s for a selection of characteristics: (a) for households in owner occupation and (b) for households renting their accommodation

Characteristic	Per cent (a) r	Per cent (b) r'	$\sqrt{\text{deff}}$ (1) All	$\sqrt{\text{deff}}$ (2) r	$\sqrt{\text{deff}}$ (3) r'	$\sqrt{\text{deff}}$ (4) $r-r'$	Corr (r, r') R
Household characteristics							
Head of household:							
is aged under 40	31.6	26.6	1.21	1.13	1.31	1.20	0.02
is chronic sick	17.1	30.9	1.31	1.16	1.13	0.97	0.28
is a hospital out-patient	10.0	12.7	1.07	1.04	1.05	1.01	0.07
has made a doctor consultation	10.2	16.0	1.03	0.93	1.07	0.99	0.05
is working	74.4	66.5	1.29	1.17	1.25	1.11	0.16
Single-person household	13.0	22.2	1.15	1.13	1.08	1.04	0.11
Household has no car	31.3	60 8	1.57	1.43	1.25	1.27	0.11
Accommodation characteristics							
Pre-1919 building	33.4	27.7	2.25	1.78	2.04	1.47	0.40
Has a fixed bath	95.3	88.7	1.92	1.35	1.83	1.37	0.41
Has an inside lavatory	92 9	82.2	2.07	1.53	2 05	1.64	0.30
Has fixed central heating	40 0	16.5	1.72	1.49	1.72	1.43	0.17
Is a whole detached house	28.1	4.2	1.91	1.72	1.13	1.57	0.17
Sample size			3346	1584	1762		

Equation (6.26) can be used in various, slightly different, ways to model sampling errors for differences.

(1) One way is to use it to estimate R from v(r), v(r') and v(r-r') computed for a large number of statistics; average these estimates over appropriately defined groups of statistics; and then use the averaged R in reverse to predict v(r-r') from v(r) and v(r') using equation (6.26).

(2) On the assumption that $deft_s$ values are similar for the two subclasses being compared, (6.26) gives

$$deft_d^2 = (1-R).deft_s^2.$$

With R estimated as in (1), the above equation may be used to predict $deft_d$ for the difference from the (averaged) $deft_s$ for the two subclasses, which can then be used to predict v(r-r') from $v_o(r)$ and $v_o(r')$ using (6.26).

(3) A particularly convenient approach is to seek for var(r-r') a model of exactly the same form as developed above for subclass variances. Assuming that population variance for the two subclasses is similar, we can write (6.26') in the form

$$\left(\frac{s^2}{n_d}.D_t^2\right) < v(r-r') < \left(\frac{s^2}{n_d}.D_t^2\right)\left[1 + 2.\frac{n_d}{n_t}.d_t^2\right] \qquad (6.27)$$

where n_d is half the harmonic mean of the two subclass sizes say n_a and n_b

$$n_d = \frac{n_a.n_b}{n_a + n_b}. \qquad (6.28)$$

D_t is the effect of weighting which is common to all estimates whether concerning the total sample, subclasses or subclass differences (assuming cross-classes); and $deft_t$ is the total sample design effect for a given variable. Here the approximation made is:

$$\frac{m_s.\bar{b}_t - 1}{\bar{b}_t - 1} \approx m_s$$

Even for large subclasses, the range implied in (6.27) is usually quite narrow. For instance for $deft_t^2 = 2$, $D_t = 1$, and $n_a = n_b = 0.2.n_t$ (each subclass forming 20% of the total sample), the range in (6.27) becomes

$$10.\frac{s^2}{n_d} < v(r-r') < 12.\frac{s^2}{n_d}.$$

Hence it appears reasonable to take v(r-r') in the middle of this range, giving

$$v(r-r') = \left(\frac{s^2}{n_d}.D_t^2\right)\left[1 + \frac{n_d}{n_t}.d_t^2\right] \qquad (6.29)$$

The remarkable thing about (6.29) is that it is exactly in the same form as (6.25) for subclasses, except for the replacement of subclass size n_s by n_d defined in (6.28). This means any model developed for individual subclasses can be extended to subclass differences as well. Table 7F.(3) in Chapter 7 provides an example of a relationship of this form presented in a tabular form. In practical applications, it is always useful to allow for some empirical determined parameter(s) to improve goodness of fit of the model with actual computations. For instance, one may modify (6.28), which defines the 'effective sample size' to be used in (6.24) or (6.25) for a subclass difference, to

$$n_d = \beta.\frac{n_a.n_b}{n_a + n_b}.$$

where ß is an empirically determined factor, expected to be close to 1.0.

<u>More Complex Statistics</u>

Subclass differences represent a basic measure of relation between variables. Empirical findings about them lead to conjectures about design effects for other statistics that measure relations, such as regression coefficients. On the basis of semi-empirical considerations, Kish and Frankel (1974) conclude the following in relation to deft for an analytical statistic, say z, such as a correlation or regression coefficient:

1. *deft*(z)>1. In general, design effects for complex statistics are greater than unity. Hence standard errors based on simple random sample assumptions tend to underestimate the standard error for complex statistics.
2. *deft*(z)<*deft*($\bar{y}$). Design effects for complex statistics tend to be less than those for means, for a given variable and sample or subclass. The latter are more easily computed and tend to provide 'safe' overestimates for the former.
3. *deft*(z) is related to *deft*($\bar{y}$). For variables with high defts for means, values of *deft(z)* also tend to be high.
4. *deft*(z) tends to resemble *deft*($\bar{y}_a-\bar{y}_b$), the design effect for differences between means.
5. *deft*(z) tends to have measurable regularities for different statistics.

Based on the above, the authors propose a simple model

$$\frac{deft^2(z) - 1}{deft^2(\bar{y}) - 1} = k$$

with *deft*($\bar{y}$)>1; and k (0 <k <1) being specific to a particular variable, type of statistic, and sample or sample subclass.

7

PRESENTATION AND USE OF INFORMATION ON SAMPLING ERRORS

As has been emphasised earlier, even when extensive computations of sampling errors can be undertaken, their presentation in a suitable form remains a problem in large-scale, multi-purpose surveys. Obviously the presentation with each and every survey estimate of its associated sampling error is out of the question, since that would double the size of the publication. Nor would such undigested presentation be useful, since results of individual computations are not always reliable, given the variability of sampling error estimates themselves.

This chapter is aimed at providing general guidelines on presentation of information on sampling errors for different types of users and uses of the survey results. Perhaps the most fruitful means to do so is to consider a number of illustrations, as presented later in this chapter. The reader's attention may be drawn to a very useful publication on the topic which also provides a number of illustrations, some of which will be referred to below. This is Gonzalez et al (1975), an earlier version of which is available as United States (1974).

7.1 SOME BASIC PRINCIPLES

Certain basic principles need to be observed in choosing the appropriate mode of presentation of information on sampling errors:

(1) Sampling errors must be presented in the context of the total survey error. The user should be made aware of the fact that sampling variability is just one, and not always the most significant, component of the total error.

(2) The mode of presentation and the degree of detail given should suit the specific needs of particular categories of users. There are several types of users interested in the survey results: the general, often non--statistical user with no special interest or expertise in survey methodology; the substantive analyst engaged in primary or secondary survey research; and the sampling statistician concerned with survey design and evaluation. Each type of user has his or her own specific requirements for information on sampling errors.

(3) The form of presentation should be convenient and relevant to the needs of particular categories of users, and should be so as to encourage proper use and interpretation of the information. A problem in many surveys is that the information on sampling errors, even when available, is not properly utilised.

(4) It should be remembered that, despite its importance, information on sampling errors is necessarily only secondary to the main, substantive results of any survey. Therefore it is important to ensure that the information on sampling errors does not clutter the presentation of substantive results of the survey. The objective of providing this information is to elucidate the limits to the reliability of the substantive results and not to obscure the results.

(5) The accuracy with which information on sampling errors is required depends on the specific uses to be made of the information. For many purposes approximate values or indication of the overall patterns and magnitudes will suffice. Also in the light of (3) and (4), it is often better to provide approximate information which is more likely to be applied than to provide exact information which is hard to use.

(6) For any category of users, it is important to summarise the information and provide sufficient explanation to facilitate correct interpretation of the results. Users not directly involved in the design and execution of the survey cannot be expected to be familiar with its details and peculiarities.

(7) The only basis for proper summarisation and concise presentation of the mass of information on sampling errors is to analyse thoroughly the pattern of results utilising empirical and theoretical models, such as of the type discussed in the previous chapter.

7.2 REQUIREMENTS OF DIFFERENT CATEGORIES OF USERS

7.2.1 THE GENERAL USER OF SURVEY RESULTS

Several categories of users may be distinguished. The first is the general user, perhaps with no special interest or expertise in survey methodology or substantive research, who is interested in using the survey results for drawing broad conclusions and taking decisions. For this type of user, the information on sampling variability should indicate the overall quality of the results of the survey and their place within the wider body of related statistical information. More specifically, it should indicate how substantively significant conclusions to be drawn from the survey may be affected by the uncertainties due to sampling variability.

The focus should be on how information on sampling errors (or indeed on any type of survey errors) affects the interpretation of substantively significant results of the survey. Sampling error should be placed in the context of total survey error, and viewed as the lower limit of that error. It should be indicated how sampling error becomes the critical component of total error for small subclasses and subclass differences, and how its magnitude determines the detail to which the survey data may be meaningfully cross-classified.

The text of a report presenting sampling error data should include a statement that defines and interprets basic terms such as 'sampling error', 'standard error' and 'confidence interval', etc. These concepts should be illustrated by numerical examples. Gonzales et al (1975) provide examples of an introductory text which may be used for this purpose. See also Illustration 7A below.

In view of the above, it is recommended that information of sampling errors in general survey reports should include at least the following explanatory material:

(i) a description of main sources of non-sampling errors, including coverage, non-response, response and processing errors;

(ii) definition and interpretation of the terms used in the presentation of sampling errors;

(iii) summary information on magnitude of sampling error for the more important estimates resulting from the survey; and

(iv) caution, and specific guidelines to the extent possible, on the limits in the degree of detail to which the survey results may be classified due to the presence of sampling errors. This is most important because of the constant pressure to classify the survey data in more and more detail.

For the general user, the most useful form of presentation probably is to accompany all important estimates discussed in the text with their respective sampling error, specially where the error may affect the substantive conclusions to be drawn from the surveys. Sampling errors may be presented in different forms, for example:

. as absolute values of the standard error (se);
. as relative values, standard error divided by the estimate (rse);
. in the form of probability or confidence intervals.

The preference between absolute and relative forms will depend upon the nature of the estimate. The same value of standard error may be applicable to a number of estimates when expressed in relative terms; as for example for aggregates that vary in size or in units of measurement. In such cases, it is economical as well as more illuminating for the reader to present relative standard errors.

However, absolute values of the standard error (se) are sometimes easier for the reader to relate to the estimate, especially in the case of proportions, percentages and rates. In any event, it is important to avoid ambiguity in presenting standard errors for percentages: clear distinction needs to be made between the absolute number of percentage points and the concept of relative error in percentage terms. For example for a percentage p = 40% and standard error se = 2%, the relative error is 5%, and should not be confused with the absolute value of the standard error (2 per cent).

The presentation of error in the form of probability intervals requires a choice of the confidence level. Some analysts prefer to give only the standard error (eg in parentheses following the estimate in the text, or as a separate column in text tables), so that the user can compute whatever multiple of standard error is appropriate for the desired confidence interval. However, in guiding the user in the interpretation of results when issues of statistical significance arise, it is more convenient to present the survey estimates directly in the form of confidence intervals. Since there is no widespread agreement on the appropriate choice of confidence interval (say, 90, 95, or 99 per cent), it is necessary

(a) to specify what confidence interval is being used, and

(b) to follow the same level throughout as far as possible in determining what is to be regarded as 'statistically significant'.

The most cannon practice is to use the 95 per cent confidence interval, ie

estimate ± 2.*(standard error)*

though there are many examples of survey reports which use ± one standard error as the interval.

It should be pointed out that to avoid comment when the observed difference is not 'statistically significant' is not always the appropriate solution: it may reduce the attention given to important results, or encourage an interpretation of 'no difference', or 'no change', when the band of uncertainty is large and important differences could be present. Furthermore, it is possible that significant results would emerge with less detailed classification of the sample; if so, attention should be drawn to this fact.

In many situations it is sufficient to provide only approximate information on the magnitude of the standard error. This would be the case, for example, when se (or relative error, rse) has similar values for a number of estimates, so that a single averaged value may suffice. Similarly, approximate values would suffice when the sampling error is unimportant with respect to the relationship being discussed.

In such situations a simple statement, such as 'relative error of these estimates is in the range 3-5 per cent...' may be included in the text, text tables or footnotes. Somewhat more detailed information may be provided by indicating different ranges of values of se or rse by different symbols, for example as follows:

.Relative standard error is under 5 per cent unless otherwise indicated.
*.Relative error 5-10 per cent is indicated by one asterisk**
*.Relative error 10-15 per cent is indicated by two asterisks***
.Relative error >15 per cent is indicated by enclosing the estimate in parentheses ().

A simpler version of this scheme has been used in some survey reports. To save space and improve readability, the text or summary tables in these reports generally do not indicate the number of sample cases on which estimates are based. As a safeguard to the reader, something like the following system may be used to indicate the range of sample size (rather than of the standard errors directly) for cells of the text tabulations:

.Sample size (cell frequency) > 50 unless indicated otherwise.
.If frequency 20-50, estimate enclosed in parentheses ().
.If frequency <20, estimate suppressed and replaced by an asterisk.*

It should be pointed out that the suppression of some data cells in a table because the sampling error is too large (ie cell size too small) is not in general a good practice. Suppressing of individual cell values prevents the user from combining categories of the table. Moreover, results which may not be statistically significant due to large sampling error may still be meaningful, for example in the fact that the estimate is 'small' rather than 'large'.

7.2.2 THE SUBSTANTIVE ANALYST

The second category is that of the substantive analyst engaged in further analysis and reporting of the results. This type of user requires more detailed information on sampling errors. He or she may wish to go beyond the text or text tables to look at the detailed tabulated data or to produce new tabulations, and will expect to find not only direct (computed) estimates of sampling errors for all major statistics, but also a general indication of the magnitude of standard error to be expected for any estimate over any category of the sample. These requirements suggest:

(i) A tabular presentation of computed sampling error estimates for all important variables for the total sample, for major sampling domains, and for a variety of subclasses and subclass differences.

(ii) A graphical or tabular presentation of approximate standard errors (or other measures of sampling error) for a number of variables as a function of subclass size.

(iii) Similar information for differences between subclasses.

The last two provide information in a summary form. It may be necessary to produce such summaries separately for different types of subclasses or for different sampling domains. The objective is to summarise results from detailed computations, smooth out random variability in the computed results, and provide a basis for extrapolation to statistics for which sampling errors have not been computed or tabulated. Comparison of the averaged or smoothed results with those actually computed provides the user with an impression of the degree of reliability of individual computations and of the goodness of fit of the smoothed results.

7.2.3 THE SAMPLING STATISTICIAN

The sampling statistician is concerned with the statistical efficiency of the design adopted, compared to alternatives which could have been adopted, or more relevantly, that might be adopted in future surveys with similar objectives. The type of information that is useful for sample design and evaluation includes:

(i) Detailed information on standard errors and their pattern of variation with subclass type and size.

(ii) Similar information on design effects.

(iii) Information on roh values to permit extrapolation across variables and across designs.

(iv) Information on the effect of specific features of the design, such as stratification, clustering of ultimate area units and of other higher stage units, departures from self-weighting, etc.

(v) More generally, information on components of the sampling error for multi-stage designs.

Examples are given in the remainder of this chapter to illustrate important aspects of presentation of sampling error information in survey reports.

7.3 ILLUSTRATIONS

ILLUSTRATION 7A AN EXAMPLE OF INTRODUCTORY STATEMENT ON SAMPLING ERRORS FOR THE GENERAL USER OF THE SURVEY RESULTS.

In the presentation of sampling errors it is important to provide a clear and concise description of meanings of the terms used, the measures presented, and how the information may be interpreted and used.

Relationship to other non-sampling errors should also be pointed out. The details to be given will of course vary depending upon the specific situation and requirements. There are many fine examples of the type of introductory statements on sampling errors which may be included in survey reports; the one presented here from a survey in Nepal is a particularly clear and concise one.

(Illustration 7A) INTRODUCTORY STATEMENT ON SAMPLING ERRORS IN A SURVEY REPORT: AN EXAMPLE.

The estimates in this report are obtained from a sample of about 6,000 women from the population of Nepal. If the survey was repeated a different sample of women would be obtained, and hence the resulting estimates would also differ. The sampling error of an estimate measures the degree to which the estimate would vary if different samples of women were taken. in other words, the sampling error measures the imprecision caused by limiting the enquiry to a sample of the population. An important advantage of probability sampling is that estimates of sampling errors can be obtained from the results of the single sample which is actually selected.

Non-sampling errors, such as mistakes in implementing the sample design, mistakes in the respondents' answers caused by misunderstanding or memory lapse and errors in recording the data are not taken into account in estimates of sampling error, although they certainly exist to some degree. For this reason the estimate of sampling error should be interpreted as a lower bound for the total error of an estimate.

The measure of sampling error used in this report is the standard error (SE). For certain important statistics in the text the estimated standard error is given in the form of a footnote indicated by one or more asterisks (*). For example, in Section 5.1 the estimated mean number of children ever born is 5.7, with standard error 0.16.

Standard errors have the following interpretation: if non-sampling errors are ignored, then in two samples out of three the true value lies within one standard error of the estimated value, and in nineteen samples out of twenty the true value lies within two standard errors of the estimated value. Accordingly an interval of +/-2 standard errors around the sample estimate nearly always contains the true value for the population. This interval is called a 95% confidence interval, and is commonly chosen as giving a range of possible values for the estimated quantity consistent with the data.

In the example above, the 95% confidence interval is 5.7 ± 2(0.16) = 5.38 to 6.02; that is, with 95% confidence the total number of children ever born in the population lies between 5.4 and 6.0.

Standard errors for the differences between pairs of estimates are also given in the text, and these are important for determining the likelihood that an observed difference is real or merely caused by sampling variation. For example, in Section 5.3 the current fertility of women whose husbands have "no education" is compared with the current fertility of women whose husbands have "some education'. For the 35-39 age group the estimated numbers of live births in the past five years were 1.2 and 0.9 respectively, giving an estimated difference of 0.3 children. As shown in the footnote, this difference has estimated standard error 0.12, and so a 95% confidence interval for the difference is 0.3 +/- 2(0.12) a 0.06 to 0.54.

In general one can be reasonably sure that a real difference exists if the 95% confidence interval does not include the value zero. In statistical terminology, the difference is then statistically significant at the 5% level. On the other hand, the term "not statistically significant" is used in the text to describe a difference with a 95% confidence interval which includes the value zero, and in such cases the observed difference in the sample is not necessarily reflecting a difference in the population.

In the example above, the 95% confidence interval does not cover zero, so there does appear to be a difference in the current fertility according to husband's education for the 35-39 age group. The interval (0.06 to 0.54) also implies that the magnitude of the mean difference cannot be estimated with precision from the survey but is unlikely to be more than half a child.

(cont.)

(Illustration 7A, cont.)

Sampling errors in the text are derived from data presented in Table .... and Table ..... The standard errors of estimates of 17 important variables for the whole population are given in Table .....

In addition to standard errors (SE), the following quantities are presented:

m mean or percentage value of the estimate.

n sample size.

DEFT the "design effect", a factor which compares the standard error of the actual clustered sample with the standard error expected if the sample had been selected by simple random sampling of individuals. That is, DEFT = SE/SR, where SR is estimated by the usual simple random sampling formula.

S the standard deviation, defined as $SR.n^{1/2}$. This is a measure of the variability between individuals, and is a characteristic of the population and not of the particular sample design.

In Table ...., values of m, n, and SE are given for the same set of variables for 12 subclasses of the population, defined by Age, Years Since Marriage, Age at Marriage, Literacy and Terai/non-Terai. The precision of estimates for these subclasses can be obtained from this table.

More derailed sampling errors can be made available on request. However, the following general statements can be inferred from the calculated standard errors and design effects.

(1) The standard errors for means based on the whole sample generally range between 1% and 5% of the mean.

(2) Many observed differences are not statistically significant when necessary demographic controls are introduced. Hence small differences should be interpreted with caution.

(3) The design effects for the whole sample are large for some variables, ranging from 1.14 to 4.19. (For example, a design effect of 4 for a variable implies that a random sample of 1/16 the size of the present clustered sample would achieve the same precision for that variable as that achieved by the current sample.) This is not unexpected since the survey design was highly clustered because of constraints on time and travel in difficult terrain. However, these high design effects are considerably reduced for estimates for subclasses, and further reduced for differences in subclass estimates, so this should not be taken as a compelling argument against cluster sampling. The data in those tables are of considerable interest for the design of future surveys in Nepal.

ILLUSTRATION 7B UP-FRONT PRESENTATION OF ERRORS ON IMPORTANT STATISTICS

In statistical reports aimed at the general user, it is important to bring to the readers' attention the magnitude of sampling error for at least the main estimates produced from the survey. A good strategy is to extract a few of the most important figures and present them up-front, preferably as an integral part of summarisation of the survey results. Relative standard error (ie error as percentage of the estimate) can be a particularly convenient form in the presentation of results for totals, means and other ratios. However, in relation to percentages or proportions care is needed, as noted earlier, to ensure that there is no confusion between errors presented in the form of absolute per cent points, or in relative terms as percentage of the estimated value.

The series of tables accompanying this illustration provides a wealth of information on relative errors on diverse topics from the Indian National Sample Survey, covering consumer expenditure, land holding, employment and labour force, assets and liabilities of households, and birth rates. These tables provide examples on the kind of important statistics which may be presented up-front along with information on their relative standard errors.
Of course the data on sampling errors presented in these examples are also of interest in their own right from a substantive point of view: very little of such information is published from surveys on different topics from developing countries. Even though the actual magnitudes of errors shown are a function of the particular sample sizes (and designs) involved, their relative magnitudes for different types of variables from the same or similar survey should be of considerable general interest to survey practitioners.

The data presented here originate from diverse sources and have been quoted here from a compilation in India (1990).

(Illustration 7B) **RELATIVE STANDARD ERRORS FOR DIVERSE ITEMS FROM THE INDIAN NATIONAL SAMPLE SURVEY. (Source: India, 1990.)**

[1] PER CAPITA MONTHLY CONSUMER EXPENDITURE (NSS 28TH ROUND)

State	Rural					Urban				
	cereals	total food	total nonfood	consumer expendi.	sample size*	cereals	total food	total nonfood	consumer expendi.	sample size**
Andhra Pradesh	1.41	1.68	6.15	2.60	667	1.67	1.90	6.11	2.96	375
Kerala	2.76	2.52	4.77	2.86	362	4.32	5.90	9.35	6.05	136
Madhaya Pradesh	1.46	1.38	3.68	1.68	709	2.05	2.15	5.78	2.92	295
Maharashtra	1.48	1.33	4.44	1.82	618	3.44	2.05	5.27	2.79	578
Punjab	1.87	1.89	4.95	2.53	359	2.62	2.95	7.22	3.86	140
Tamil Nadu	1.75	1.71	4.24	1.85	530	1.67	2.12	7.14	3.33	411
Uttar Pradesh	1.07	1.03	3.39	1.32	1030	1.62	1.77	4.67	2.40	578
West Bengal	1.58	1.88	4.94	2.27	540	1.99	2.81	4.79	2.66	222

*villages; ** urban blocks;
the design was self-weighting with an average of 2 households per village/block.

[2] ALL-INDIA ESTIMATES OF TOTAL NUMBER RURAL HOUSEHOLDS, TOTAL AREA OPERATED AND TOTAL AREA OWNED. (NSS 8TH ROUND)

	households	area operated	area owned	--sample size-- hhs	villages
Central Sample	1.31	2.17	2.32	24366	1410
State Sample*	0.80	1.49	1.57	47432	2805

*The State Sample is the part of the sample surveyed by State Governments.

ILLUSTRATION 7B continued:

[3] PROPORTION OF THE POPULATION (AGED 5+) EMPLOYED, UNEMPLOYED, AND NOT IN THE LABOUR FORCE (NSS 27TH ROUND)

	employed	unemployed	not in labour force	-----sample size-- hhs	villages/ blocks
rural	0.43	5.55	0.45	72,000	9,000
urban	0.47	2.93	0.28	53,000	4,800

[4] BIRTH RATE IN RURAL AREAS IN FIVE STATES (NSS 19TH ROUND)

	sample size villages*	relative standard error (%)
Andhra Pradesh	336	2.77
Bihar	384	2.57
Maharashtra	348	3.44
Rajasthan	192	3.64
Uttar Pradesh	525	2.03

*An average of approximately 20 household selected per village.

[5] VARIABLES RELATING TO ASSETS AND LIABILITIES, ESTIMATED AT ALL-INDIA LEVEL
(NSS 37TH ROUND, DEBT AND INVESTMENT SURVEY)

Item		--relative error-- rural	urban
Proportion of households			
reporting cash dues payable		1.98	3.48
cash dues payable (amount)		4.41	6.15
amount of assets		1.02	2.19
amount of land owned	1.22	3.28	
whether owns residential building		1.25	2.68
whether owns livestock		1.27	6.21
whether owns agricultural implements	2.38	10.88	
whether owns non-farm equipment		7.76	10.34
whether owns transport equipment		3.74	5.54
whether owns durable assets		1.41	1.91
all shares owned		5.73	15.63
other financial assets owned		5.20	4.37
Sample size (number of households)		61,157	30,965

ILLUSTRATION 7C GRAPHICAL PRESENTATION OF CONFIDENCE INTERVALS

Graphical presentation of sampling error information is sometimes helpful: it can convey to the reader more directly the magnitude of the degree of uncertainty and its impact on the estimates produced from the survey. Perhaps the most suitable form is to show the errors in the form of confidence intervals to some appropriate level of confidence. Illustration 7C.(1) provides an example, quoted from Gonzalez et al (1975). In comparison across subgroups in the top diagram (labelled 'A'), it is instructive to note for instance that for certain groups the confidence intervals overlap, indicating that the direction of the difference between them cannot be ascertained with the specified level of confidence. The lower part ('B') shows more directly confidence intervals of comparisons. Differences statistically not significant clearly stand out.

This type of presentation can be particularly appropriate in executive summaries or other short reports prepared for special audiences, and generally in situations when it is important to ensure that wrong impressions formed from (statistically) small differences are avoided.

The graphical presentation in Illustration 7C.(2) has a rather different objective; it is to present a relationship (such as between variance of individual subclasses and that of their difference) in a graphical form so that the user can read off some required statistic directly from the graph without going through elaborate tables of calculations. Sometimes this form of presentation can be convenient, but perhaps less often than assumed. The kind of user who is interested in obtaining information on sampling errors for estimates for which it cannot be presented in the report directly, is usually quite capable of reading off and interpolating information from tables or evaluating algebraic expressions to obtain numerical results.

ILLUSTRATION 7C.(1). GRAPHICAL PRESENTATION OF CONFIDENCE INTERVALS.

(Source: Gonzalez et al, 1975.)

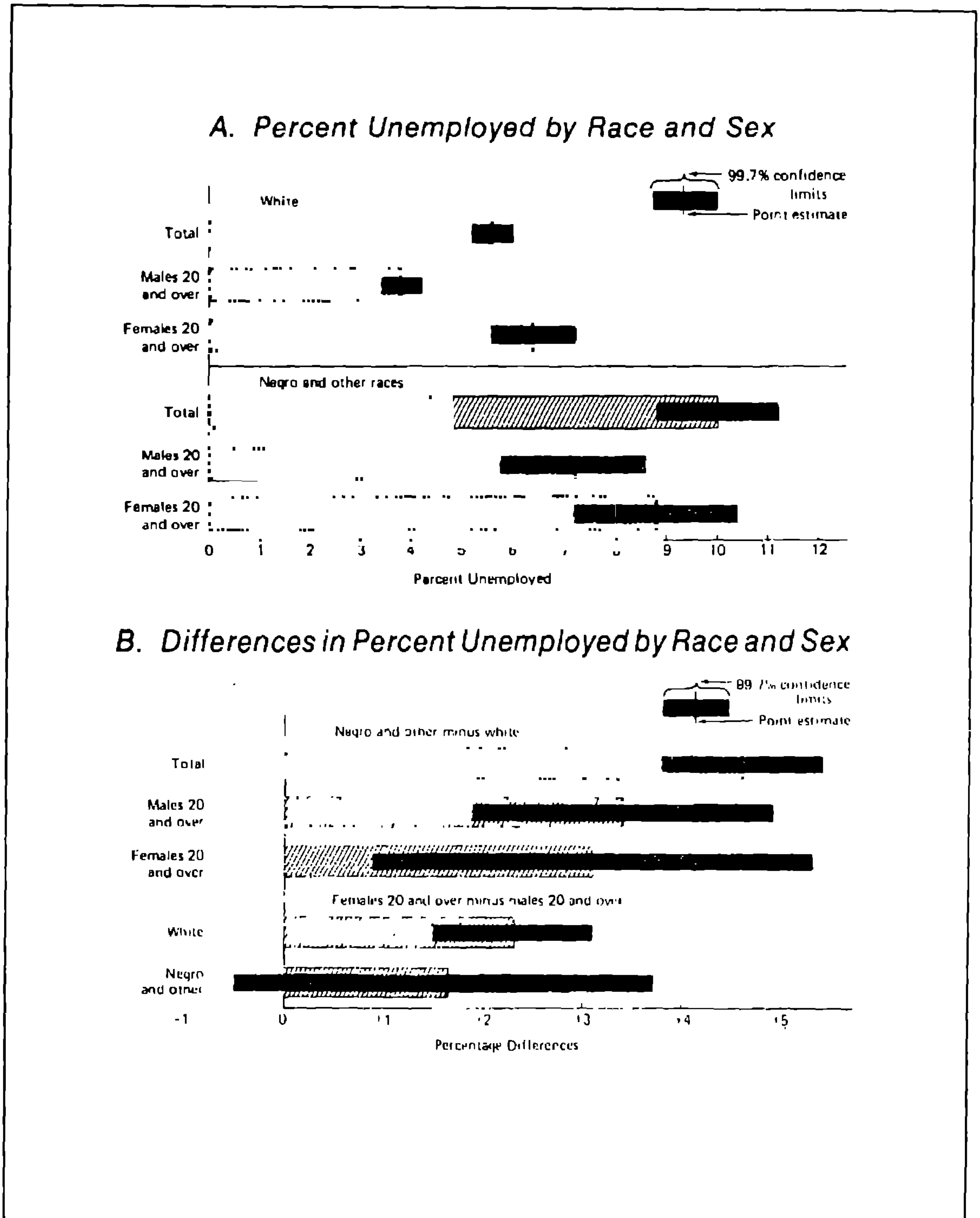
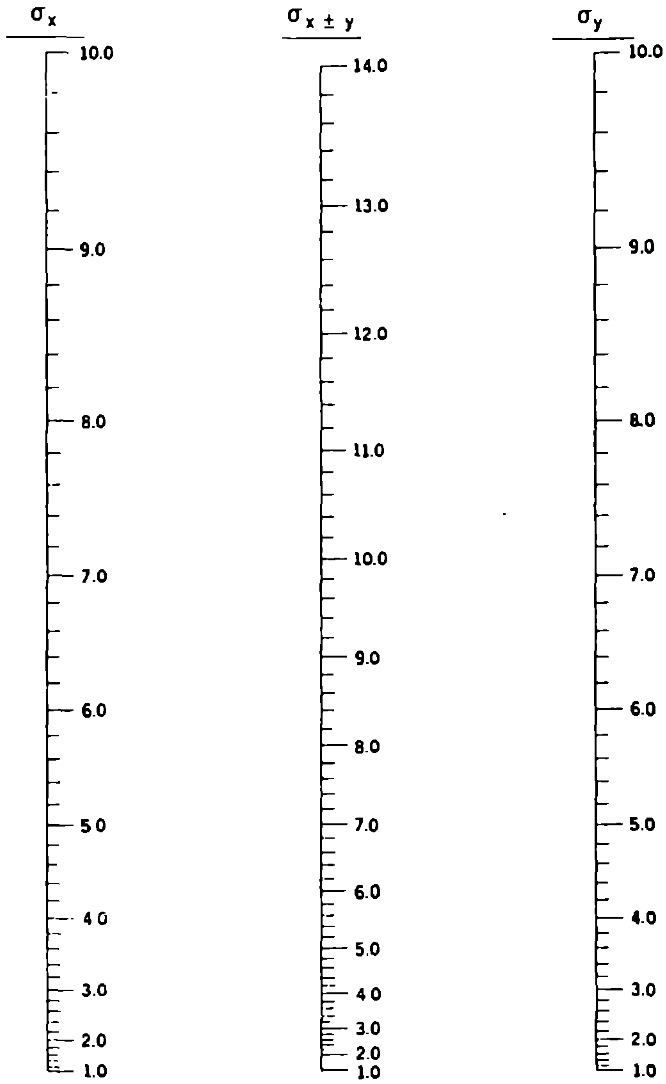


ILLUSTRATION 7C.(2).GRAPHICAL PRESENTATION OF ERRORS OF SUMS AND DIFFERENCES.

C. Nomogram: Standard Error of Sum or Difference
Independent Samples



Instructions for Use: If x and y are two independent estimates, then x + y and x − y are estimates of the sum and the difference, respectively. The standard errors may be approximated by the use of this nomogram. Locate the point on the σ_x scale that corresponds to the standard error of x, and the point on the σ_y scale that corresponds to the standard error of y. The scales may be read in any units (tenths, thousands, millions, etc.) provided that the *same* unit is used on all scales. Now connect the points by a straight line (a stretched thread is convenient) and read the value where the line crosses the $\sigma_{x \pm y}$ scale. This is the standard error of x + y and x − y. For example, suppose the standard error of x is 6,750 and the standard error of y is 4,700. A straight line between these values on the σ_x and σ_y scales crosses the $\sigma_{x \pm y}$ scale at about 8,250. An exact computation would have yielded the value 8,225.

ILLUSTRATION 7D DISPLAYING THE MARGINS OF UNCERTAINTY WITH RESULTS FROM INDEPENDENT REPLICATIONS

A powerful way of conveying to the user an impression of the margins of uncertainty in sample survey results is to display side by side estimates from a number of independent replications into which the total sample has been divided - assuming that the sample design and implementation permits the construction of independent replicated estimates.

This method has been discussed in detail in Section 3.2, where reference to the present illustration was also made. This form of presentation of course requires that the number of replications is small enough (or a larger number of replications has been appropriately collapsed into a suitable number) for the full results from each to be displayed in the survey report. In practice one would also limit such presentations to a selected subset of the most important estimates. Some authors have criticised this form of presentation because of the usually very small number of replications involved, and the possible lack of independence among the replicated estimates. The method may also not be considered practical for multi-purpose surveys involving many variables and forms of analysis. Nevertheless, this form of display of the margins of uncertainty is generally useful and instructive - provided of course the basic assumptions underlying the method are satisfied in the survey design and implementation.

The illustration below is from an earlier round of the Indian National Sample Survey. (It is due to Mahalanobis, quoted in Zarchovich 1965, Chapter 8.) The total sample for the NSS Round 8 (1954-55) was enumerated in two parts: one administered centrally and the other at the level of state. In the illustration the former part has been divided into four independent replications, and the latter into eight. A set of results on variability between the sample results on the numbers of holdings by size is shown, both between the central and state parts, and between subsamples within each part.

With independent implementation of data collection and processing by subsample, the observed variability reflects the effect of sampling as well as several sources of non-sampling variation. Actually in the NSS, the surveys go beyond the concept of simple independent replication in design and implementation. Rather, the more comprehensive concept of 'interpenetrating' subsamples which can be linked between as well as within PSUs is used to control and study the reliability of survey results more comprehensively (Lahiri, 1958). As noted in <u>A Dictionary of Statistical Terms</u> (Kendall and Buckland, fourth edition 1982), interpenetrating sampling refers to the taking of two or more samples from the same population by the same process of selection, but "the sample may or may not be drawn independently, linked interpenetrating sampling being an example of the latter". Also, "there may be different levels of interpenetration corresponding to different stages of a multistage sampling scheme... Generally the subsamples are distinguished not merely by the act of separation into subsamples but by definite differences in survey or processing features, eg when different parts are assigned to different subsamples, or one subsample is taken earlier in time than others."

Note that the separate results from independent replications can provide only a rough indication of the conventional measures of standard error and confidence intervals.

(Illustration 7D) DISPLAY OF ESTIMATIONS FROM INDEPENDENT REPLICATIONS: NSS, INDIA.

TABLE 30. - ESTIMATED NUMBER OF HOUSEHOLDS BY SIZE CLASSES OF AREA OWNED [1]
(National Sample Survey, 8th round, July 1954-April 1955, rural sector, All-India)

	Holding size [2] (acres)	Central sample					State sample									Central and state samples combined
		s.s. 1	s.s. 2	s.s. 3	s.s. 4	Combined	s.s. 1	s.s. 2	s.s. 3	s.s. 4	s.s. 5	s.s. 6	s.s. 7	s.s. 8	Combined	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1	0.00 [3]	13 855	15 036	15 617	13 538	14 511	13 874	13 791	15 140	14 819	16 380	15 474	13 315	14 811	14 699	14 641
2	0.01-0.04	6 901	5 948	6 505	6 814	6 541	5 444	5 165	4 893	4 780	5 701	4 794	5 150	4 748	5 085	5 570
3	0.05-0.09	1 904	1 699	1 717	1 690	1 752	1 726	1 861	1 856	1 281	1 559	1 253	1 992	1 527	1 632	1 672
4	0.10-0.49	3 779	4 258	3 983	3 938	3 990	4 334	4 487	5 081	4 342	3 801	4 299	3 689	3 842	4 234	4 153
5	0.50-0.99	4 136	4 072	3 947	3 752	3 976	3 667	4 146	4 294	3 963	3 796	3 460	4 120	4 078	3 940	3 953
6	1.00-1.49	3 566	3 390	3 739	3 483	3 544	3 557	3 566	3 526	3 220	3 926	3 390	3 427	3 314	3 492	3 509
7	1.50-2.49	5 667	5 599	5 627	4 701	5 400	4 925	5 218	5 097	5 480	5 592	5 431	5 473	5 669	5 360	5 374
8	2.50-4.99	8 915	9 033	8 630	8 795	8 841	8 323	8 401	8 758	8 935	8 509	8 625	7 751	7 998	8 412	8 557
9	5.00-7.49	5 904	4 821	5 585	4 709	5 256	4 806	5 072	4 931	4 432	4 497	5 100	4 884	5 156	4 861	4 992
10	7.50-9.99	3 395	2 783	2 741	2 919	2 961	3 118	2 955	2 980	2 768	3 072	3 042	3 135	3 036	3 013	2 995
11	10.00-14.99	3 431	3 289	2 778	3 090	3 146	3 381	3 060	3 446	3 607	3 099	3 276	3 083	3 191	3 267	3 227
12	15.00-19.99	1 700	1 784	1 742	1 792	1 757	1 664	1 505	1 567	1 718	1 517	1 730	1 949	1 526	1 648	1 682
13	20.00-24.99	846	1 036	795	932	901	858	907	1 009	959	1 121	809	964	942	945	931
14	25.00-29.99	713	691	637	607	659	570	541	495	645	659	635	753	683	622	636
15	30.00-39.99	644	761	739	615	692	568	657	766	712	724	776	715	750	709	702
16	40.00-49.99	317	341	356	405	354	355	330	349	410	299	395	410	314	358	357
17	50.00-74.99	352	369	368	366	363	411	321	292	312	404	471	424	387	378	373
18	75.00-99.99	68	83	141	123	104	69	83	105	195	131	118	92	91	110	108
19	100.00-249.99	81	122	89	108	101	100	131	93	144	110	113	74	151	116	109
20	250.00-499.99	11	12	8	6	10	5	7	—	7	16	19	3	14	10	8
21	500.00 and above	—	7	8	--	4	3	--	—	--	3	3	--	4	1	3
22	Total	66 185	65 134	65 752	62 383	64 863	61 758	62 204	64 678	62 729	64 916	63 213	61 403	62 232	62 892	63 552
23	Number of sample villages	353	351	353	353	1 410	355	356	355	356	345	347	347	344	2 805	4 215
24	Number of sample households	6 118	6 362	6 068	5 818	24 366	6 017	5 638	5 936	6 105	6 077	6 199	5 700	5 760	47 432	71 798

[1] This table is based on the results which were presented by Professor P.C. Mahalanobis in his lecture at the UN/FAO Far East Training Center on Census Techniques, held at Tokyo, 1958.
[2] Reference period: major crop season, 1953/54.
[3] Households owning either no land or land less than 0.005 acre are included in this size class.

ILLUSTRATION 7E CONCISE SUMMARIES OF STANDARD ERRORS FOR NUMEROUS PROPORTIONS AND COUNTS IN LARGESCALE SURVEYS.

This illustration is similar to Illustration 6B in Chapter 6, and provides a more recent example of how a vast quantity of information on sampling errors may be summarised and presented in a concise manner. The main limitation of the form of presentation is that it is basically suited only to estimates of counts or proportions; it is more difficult to deal with means and aggregate values of substantive variables. Nevertheless there are many censuses and surveys where the statistics of interest are primarily in the form of counts or proportions.

The United States 1980 census of population involved the enumeration of a small number of items on a 100% basis, supplemented by an approximately 20% sample of housing units and persons for more detailed information. To expedite data release, the census sample itself was subsampled to obtain an Early Release Sample (ERS, see United States, 1982). The ERS involved basically a two stage design: PPS systematic sampling of census EDs followed by the selection of housing units or persons. Roughly the ERS consisted of 17,000 EDs of average size around 70 persons, from which an average of 20 persons were selected per ED; the overall sampling rate was around 1 in 62. Sampling errors were computed for 1,120 separate estimators for each of the 89 'publication areas' into which the country had been divided. Sampling error information was presented primarily in the form of design effects (deft). To reduce the amount of information to be presented, a two step data reduction procedure was employed.

(1) The 1,120 data items were aggregated into 36 groups based on subjective judgement about similarity of defts For easy reference and use, the groups also had to be substantively homogeneous. In each publication area separately, a weighted average of computed defts for individual data items was taken. Defts rather than ratios of variances ($deft^2$) were averaged because the former are less affected by extreme values. The weights were in proportion to the total estimated count for the data item concerned.

(2) Next, the 89 publication areas were aggregated into 10 groups using a clustering procedure based on deft values for certain selected population characteristics relating to education, employment and income. The averaged defts from (1) were themselves averaged over these groups, with weights in proportion to publication area population.

These averaged values are shown in Table 7E.(2). For any particular data item in a particular publication area, the deft is taken to be the averaged value for its data/publication area group. Standard error for the item is obtained by multiplying this deft value by the simple random sampling error from Table 7E.(1). This is simply a tabulation of the standard error as a function of the population base (N) of a 'publication area group', and estimated proportion (p) or count (Y=p.N) of units with any specified characteristics:

$$se(p) = \left[\frac{p.(1-p)}{n}\right]^{1/2} = [62.p(1-p)/N]^{1/2}$$

$$se(Y) = N.se(p) = \left[62.Y.(1-\frac{Y}{N})\right]^{1/2}$$

In using Table 7E.(1), the actual item or publication area group is not relevant except for its size. It is this feature, and the grouping of the publication areas in Table 7E(2), which make this form of presentation so concise.

ILLUSTRATION 7E.(1). STANDARD ERRORS OF PROPORTIONS AND COUNTS ASSUMING SIMPLE RANDOM SAMPLING.

(Source: United States, 1982.)

Unadjusted Standard Errors for Estimated Totals

Estimated 1/ Total	Size of Publication Area 2/						
	100,000	500,000	1,000,000	5,000,000	10,000,000	25,000,000	United States
1,000	250	250	250	250	250	250	250
2,500	390	390	390	390	390	390	390
5,000	540	550	550	560	560	560	560
10,000	750	780	780	790	790	790	790
25,000	1,080	1,210	1,230	1,240	1,240	1,240	1,240
50,000	1,240	1,670	1,720	1,750	1,760	1,760	1,760
100,000	-	2,230	2,360	2,460	2,480	2,480	2,490
250,000	-	2,780	3,410	3,840	3,890	3,920	3,930
500,000	-	-	3,940	5,280	5,430	5,510	5,560
1,000,000	-	-	-	7,040	7,470	7,710	7,860
5,000,000	-	-	-	-	12,450	15,750	17,410
10,000,000	-	-	-	-	-	19,290	24,350

1/ For estimated totals larger than 10,000,000 the standard error is somewhat larger than the table values. The formula given below should be used to calculate the standard error.

2/ Total count of persons, families, households or housing units in an SMSA, State or the U. S.

$$Se(\hat{Y}) = \sqrt{62\ \hat{Y}\ (1 - \frac{\hat{Y}}{N})}$$

N = Size of area

$\hat{Y}$ = Estimate of characteristic total

Note: Illustration 7E.(1) continued on next page.

ILLUSTRATION 7E.(1) CONTINUED:

Unadjusted Standard Errors for Estimated Percentages
(in Percentage Points)

Estimated Percentage	Base of Percentage[1]							
	5,000	7,500	10,000	25,000	50,000	100,000	250,000	500,000
2 or 98	1.6	1.3	1.1	0.7	0.5	0.3	0.2	0.2
5 or 95	2.4	2.0	1.7	1.1	0.8	0.5	0.3	0.2
10 or 90	3.3	2.7	2.4	1.5	1.1	0.7	0.5	0.3
15 or 85	4.0	3.2	2.8	1.8	1.3	0.9	0.6	0.4
20 or 80	4.5	3.6	3.1	2.0	1.4	1.0	0.6	0.4
25 or 75	4.8	3.9	3.4	2.2	1.5	1.1	0.7	0.5
30 or 60	5.1	4.2	3.6	2.3	1.6	1.1	0.7	0.5
35 or 65	5.3	4.3	3.8	2.4	1.7	1.2	0.8	0.5
50	5.6	4.5	3.9	2.5	1.8	1.2	0.8	0.6

1/ For a percentage and/or a base of percentage not shown in the table, the formula given below may be used to calculate the standard error.

$$Se(\hat{p}) = \sqrt{\frac{62}{B}\,\hat{p}\,(100-\hat{p})}$$

B = Base of Estimated Percentage

$\hat{p}$ = Estimated Percentage

ILLUSTRATION 7E.(2). DESIGN EFFECTS (DEFTS) TO BE APPLIED TO 7E.(1).

Factor to be Applied to Unadjusted Standard Errors by Characteristic and Publication Area Group

Characteristic	Publication Area Group										
	1	2	3	4	5	6	7	8	9	10	11
PERSON CHARACTERISTICS											
Urban and Rural	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Age and Sex	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Household Type, Size and Relationship	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Family Type	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Marital Status	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Nativity and Place of Birth	6.4	10.6	9.2	7.0	7.2	6.0	6.3	6.4	4.8	3.8	2.6
Language Spoken at Home and Ability to Speak English	6.0	10.1	8.9	7.8	4.7	4.1	4.2	4.4	3.0	3.1	2.1
Means of Transportation to Work	4.3	8.2	6.5	5.0	4.5	3.6	4.0	4.1	3.2	2.6	1.7
School Enrollment	3.1	4.9	4.4	3.7	3.2	2.6	2.9	2.8	2.3	2.1	1.4
Years of School Completed	4.6	8.2	6.8	5.5	4.8	4.2	4.1	4.5	3.4	2.8	1.7
Labor Force Status Excluding Unemployment	5.7	11.4	8.9	6.8	5.8	5.3	3.9	6.0	4.2	3.1	1.8
Veteran Status	2.9	4.7	4.5	3.3	2.6	2.4	2.4	2.5	1.9	1.7	1.1
Work or Public Transportation Disability Status	2.8	4.5	3.9	3.3	3.0	2.3	2.5	2.5	2.0	1.7	1.2
Unemployment	2.7	4.5	3.8	3.2	2.8	2.1	2.4	2.4	2.0	1.8	1.3
Occupation and Industry	2.3	5.4	4.7	3.8	3.3	2.7	2.8	3.2	2.4	2.1	1.4
Weeks Worked in 1979	3.0	5.0	4.2	3.4	3.1	2.4	2.5	2.8	2.2	1.9	1.3
Workers in Family in 1979	3.1	5.0	4.3	3.6	3.2	2.5	2.8	3.0	2.2	2.0	1.3
Family or Household Income	2.6	4.3	3.7	3.1	2.8	2.3	2.4	2.6	1.9	1.8	1.1
Unrelated Individual Income	4.3	6.0	5.8	5.0	4.3	3.2	4.4	4.3	3.1	2.7	1.9
Poverty Status in 1979 for Persons	6.5	11.8	9.9	7.7	6.7	5.9	5.5	5.2	4.6	3.6	2.4
Poverty Status for Families	3.6	6.7	5.6	4.0	3.7	3.3	3.2	2.9	2.5	1.9	1.2
Black or Spanish Labor Force	8.5	13.4	12.7	10.2	8.6	7.1	7.5	7.5	5.7	4.5	3.3
Black or Spanish Income	3.5	5.4	5.2	4.2	3.6	2.9	3.4	3.0	2.4	1.9	1.5
Black or Spanish Poverty	10.8	16.5	14.8	12.1	11.5	9.0	10.0	8.4	7.6	5.2	3.8
HOUSING CHARACTERISTICS											
Tenure and Vacancy Status	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Year Householder Moved Into Unit	3.8	6.2	5.4	4.7	3.9	3.2	3.4	3.8	2.8	2.6	1.7
Bathrooms	6.3	11.0	9.3	7.8	6.7	5.3	6.0	6.4	4.6	4.3	2.7
Kitchen Facilities	5.7	11.5	9.3	4.8	3.9	2.8	3.0	4.4	2.8	2.1	2.1
Sewage System and Source of Water	8.2	10.4	7.9	10.6	10.0	7.2	7.7	7.4	6.0	5.5	3.7
Air Conditioning, Heating Equipment and Utilities	6.4	12.7	9.8	8.4	7.7	5.9	6.3	6.3	5.2	4.7	3.1
Units in Structure	6.7	13.7	10.5	8.1	7.0	5.7	6.2	6.4	4.8	4.5	2.7
Year Structure Built	5.7	10.9	8.9	6.6	6.3	4.6	5.3	5.2	4.1	4.0	2.4
Bedrooms	4.7	7.8	6.8	5.9	5.0	3.8	4.1	4.6	3.4	3.2	1.9
Stories in Structure and Passenger Elevator	11.7	14.0	13.2	10.2	8.6	7.1	7.9	5.2	6.0	5.6	4.0
Mortgage Status and Selected Owner Costs	3.0	4.8	3.9	3.6	3.4	2.6	2.9	2.9	2.3	2.1	1.4
Gross Rent	4.3	6.6	6.1	4.9	4.2	3.5	3.8	3.8	3.1	2.8	1.7

ILLUSTRATION 7F CONCISE PRESENTATION OF SAMPLING ERRORS FOR DIVERSE SUBCLASSES AND SUBCLASS DIFFERENCES.

This illustration is discussed in some detail as it shows how in a multisubject survey with a complex sample design, the great volume of information on sampling errors for diverse variables, subclasses and subclass differences can be summarised for presentation. This example also represents a most carefully worked out form of presentation from a survey in a developing country. The illustration is taken from the 1976 Indonesia Fertility Survey (Indonesia 1978).

The Indonesian survey covered the six provinces in Java and Bali accounting for around two-thirds of the national population. Basically, the sample was selected in two stages: selection of around 400 clusters of average size around 100 households each with systematic PPS; followed by the selection of an average of 25 households per cluster with inverse PPS. The final sample consisted of just over 9000 ever-married women in the childbearing ages who were interviewed in detail on fertility and related factors and formed the main units of analysis for the survey. The sampling rates varied significantly across geographical domains (urban and rural areas, and provinces) but the sample of women was essentially selfweighting within each domain.

An idea about the extensive set of sampling error computations is provided by the following figures, which are not untypical for a national household survey of this type. Errors were computed for 25 substantive variables, each over the whole sample, 22 demographic and socioeconomic subclasses and 11 differences between pairs of subclasses; the whole set was repeated over 7 geographical domains, thus involving a total of nearly 6,000 estimators and their sampling errors. The set of computations could be undertaken without a great difficulty using the CLUSTERS program described in Chapter 4.

Overall results for the 25 variables over the full sample are shown in Table 7F.(1). The variables can be divided into four substantive groups covering marriage, fertility, fertility preferences and contraception. Such grouping can be helpful in the search for regular patterns in the variation of sampling errors. The table shows standard errors and some derived measures including deft. The last column shows the average cluster size. This can be used with deft (and after removing the effect of sample weights as explained in Section 6.5) to compute rohs. Also it can be useful to add to the table quantities like relative error (rse = se/r) and coefficient of variation (cv = s/r) for means.

Illustration 7F.(1) Overall results (total sample).

Variable	r	SE	95% CON. INT.		n	s	DEFT	$\bar{b}$
			r-2SE	r+2SE				
1. Age at Marriage	15.3	0.060	15.2	15.5	6341	3.26	1.54	16.9
2. First Marriage Dissolved	0.40	0.007	0.39	0.42	9136	0.47	1.44	24.3
3 Remarried	0.78	0.011	0.76	0.80	3253	0.42	1.50	8.7
4 Exposed	0.62	0.007	0.60	0.63	9136	0.49	1.36	24.3
5. Children Ever Born	3.46	0.039	3.39	3.54	9136	2.78	1.34	24.3
6 Births in First 5 Years	1.24	0.016	1.21	1.27	7428	0.95	1.38	19.8
7. First Birth Interval	24.1	0.250	23.6	24.6	6869	14.2	1.45	18.4
8. Births in Past 5 Years	0.98	0.019	0.94	1.01	6975	0.94	1.54	16.2
9. Closed Birth Interval	40.3	0.480	39.3	41.2	6660	25.6	1.54	17.7
10 Open Birth Interval	43.7	1.100	41.5	45.9	5193	50.7	1.56	13.8
11 Months Breast-fed	16.8	0.150	16.4	17.1	1933	6.20	1.09	5.1
12 Pregnant	0.10	0.004	0.09	0.11	9136	0.29	1.31	24.3
13 Wants No More Children	0.39	0.010	0.37	0.41	6744	0.48	1.71	17.9
14 Prefers Boy	0.35	0.011	0.33	0.37	3339	0.48	1.34	8.9
15 Last Child Unwanted	0.17	0.006	0.16	0.18	8218	0.38	1.44	21.9
16 Additional Number Wanted	1.05	0.035	0.98	1.13	6002	1.67	1.62	16.0
17 Desired Family Size	4.12	0.041	4.04	4.21	8681	2.03	1.88	23.1
18. Knows Modern Method	0.75	0.010	0.73	0.77	9136	0.43	2.24	24.3
19 Ever Used Pill	0.23	0.009	0.21	0.25	9136	0.44	1.95	24.3
20 Ever Used IUD	0.07	0.005	0.06	0.08	9136	0.25	1.90	24.3
21. Used Any Method	0.34	0.009	0.33	0.36	9136	0.48	1.79	24.3
22. Used Modern Method	0.30	0.009	0.28	0.32	9136	0.47	1.82	24.3
23. Using a Folk Method	0.02	0.002	0.01	0.02	5778	0.11	1.32	15.4
24 Using Any Method	0.37	0.011	0.35	0.39	5778	0.46	1.67	15.4
25. Contracepting and Wanting No More Children	0.53	0.015	0.50	0.56	2393	0.49	1.48	6.4

Notes

r = Sample estimate of ratio, mean, or proportion.
SE = Standard error of r, for the clustered sample
95% CON INT. = The 95 confidence interval, r ± 2SE.
n = Unweighted sample size.
s = Standard deviation.

DEFT = Design effect = $SE / \frac{s}{\sqrt{n}}$

$\bar{b}$ = Average unweighted number of individuals per sample PSU

A very concise form of presentation of sampling errors for diverse subclasses is possible when standard errors and defts for each variable (or group of variables) of interest can be shown as a function only of subclass size, irrespective of the nature of the characteristic defining the subclass (see Section 6.6). This is shown in Table 7F.(2) for standard errors. The table provides a good approximation for cross-classes such as age groups, and also classes which are found in most clusters but are not distributed so uniformly. It is important to recognise some limitations of such a presentation. The table does not apply as such to highly segregated or geographical classes. Also for certain variables highly correlated to characteristics defining a subclass, correction is required to account for differences in population variances of the variable in different subclasses. This is done in a very approximate manner in the footnotes following the table. Also, separate tables like this are required for each geographical domain in principle.

A similar table can be constructed for defts as a function of subclass size. (This is not shown here, but see Illustration 6G). Such a table is in fact easier to construct for defts than for standard errors, because the former measure is more portable.

Table 7F.(3) shows how Table 7F.(2) for subclasses can also be used for subclass differences. This is constructed on the basis of Equation (6.29) discussed in the previous chapter.

In the present example, the sample data are weighted, because the sample is self-weighting only within but not across domains. The weighting of sample data has important consequences for the presentation of sampling errors. Firstly, the modelling of sampling errors has to take into account the loss factors D (D_t for the total sample, which usually exceeds the factors D_g for individual domains) as explained in Section 6.6.3. Secondly, with weighted samples, special provision is usually required to provide the user with unweighted sample sizes to be used in tables like 7F.(2) and 7F.(3). For selfweighting samples this presents no problem if the good practice of showing relevant sample sizes in the tabulations is followed. However, in weighted samples it is generally not convenient to show both weighted and unweighted frequencies in the tabulations, and preference is given to showing the former. This is because exact weighted frequencies are required to permit amalgamation of categories in the tables, while unweighted frequencies are required only approximately as an indicator of (and to serve as a basis for estimating) sampling error. Consequently, additional information is required to convert the weighted frequencies shown in survey tabulations to the actual sample size, even if approximately. This is the purpose of Table 7F.(4).

Another major issue in concise presentation is the need to provide sampling errors for separate geographical domains. One option is to repeat the above form of presentation for each domain. This may be possible and desirable for major divisions of the sample, such as into urban and rural sectors. However, for smaller and more numerous domains, more concise (but necessarily more approximate) procedures may be preferable. This is not included in the presentation, but a possible approach was discussed in the context of Illustration 6G.(1) in the last chapter.

ILLUSTRATION 7F.(2). STANDARD ERRORS AS FUNCTION OF SUBCLASS SIZE.

Variable	Unweighted subclass size (n_s)											
	30-50	51-100	101-200	201-400	401-700	701-1000	1001 1500	1501-2000	2001-3000	3001-5000	5001 7000	>7000
1 Age at Marriage	0 53	0 40	0 30	0 22	0 17	0 14	0.12	0 11	0.09	0 08	0.07	0 06
2 First Marriage Dissolved	.080	060	.045	.030	.025	.020	017	.014	013	011	008	007
3. Remarried	065	050	035	.025	020	017	015	.013	.012	009	008	.007
4. Exposed	080	060	040	030	.022	018	016	.013	012	010	.008	007
5. Children Ever Born[1]	.450	340	.240	180	130	110	.090	.080	.070	060	.050	040
6 Births in First 5 years	.160	120	.090	.060	050	.040	.035	.030	.025	.020	.017	.015
7 First Birth Interval	2 35	1 80	1 30	0 95	0 72	0 60	0 50	0 44	0 38	0 31	0 26	0 21
8 Births in Past 5 years	.165	125	.090	.065	050	042	.035	.031	.027	.022	.018	.017
9 Closed Birth Interval[2]	4 35-	3.25	2 35	1 75	1 35	1 10	0 95	0 80	0 65	0 60	0.50	0.45
10 Open Birth Interval[2]	8 35	5 90	4 70	3 50	2.70	2.25	1 90	1.65	1 45	1.20	1.05	0.92
11 Months Breast-fed	0 98	0.78	0 52	0 37	0 28	0 23	0 20	0 17	0 14	0 12	0.10	0 09
12 Pregnant	045	032	.024	018	013	.011	.009	.008	.007	.006	.005	004
13 Wants No More Children	.080	.060	.045	.035	026	.022	.020	.016	.015	.013	.010	.009
14. Prefers Boy	.080	060	.045	032	024	.020	.018	.014	.013	.011	.009	007
15 Last Child Unwanted[3]	060	045	.035	.025	019	.016	.013	.011	.010	008	.007	006
16. Additional Number Wanted[4]	.265	.190	.140	.105	080	.070	060	.050	.045	037	.032	028
17 Desired Family Size	335	260	.195	145	115	100	.085	075	065	055	.045	.040
18 Knows Modern Method	075	060	045	030	025	020	.018	.015	.014	.012	.010	.009
19. Ever Used Pill	.065	050	.040	030	022	.019	.017	.014	.013	011	.009	008
20 Ever Used IUD	040	030	025	018	014	012	009	.008	007	007	006	005
21. Used Any Method	080	.060	.045	.035	025	.022	.020	.016	015	013	.010	.009
22 Used Modern Method	080	060	.045	035	025	.022	.020	016	015	013	.010	009
23. Using Folk Method	.025	020	.014	010	008	.006	.005	.004	003	.003	.002	002
24 Using Any Method	080	060	.045	035	025	022	.020	016	015	013	010	009
25 Contracepting and Wanting No More Children	080	060	045	035	025	022	.020	.016	015	013	.010	009

Notes
[1] For subclasses with mean <2 5, multiply shown value of SE by *0 5*
[2] For variables '9' and '10', multiply shown value by *0 7* for subclasses with mean <40 0, and multiply shown values by *1 3* for subclasses with mean >45 0
[3] For subclasses with proportion <0 1, multiply shown values of SE by *0 5*
[4] For subclasses with mean <0 5, multiply shown values of SE by *0 5*

ILLUSTRATION 7F.(3). EFFECTIVE SAMPLE SIZE FOR SUBCLASS DIFFERENCES AS FUNCTIONS OF SUBCLASS SIZES (n_1 and n_2).

	n_1 ($<n_2$) 100	200	400	600	1000	1500	2000	2500	3000	4000	5000
100	50	—	—	—	—	—	—	—	—	—	
200	70	100	—	—	—	—	—	—	—	—	—
400	80	130	200	—	—	—	—	—	—	—	—
600	90	150	240	300	—	—	—	—	—	—	—
1000	90	170	290	380	500	—	—	—	—	—	—
n_2 1500	90	180	320	430	600	750	—	—	—	—	—
2000	100	180	330	460	670	860	1000	—	—	—	—
2500	100	190	340	480	710	940	1110	1250	—	—	—
3000	100	190	350	500	750	1000	1200	1350	1500	—	—
4000	100	190	360	520	800	1090	1330	1540	1710	2000	
5000	100	190	370	540	830	1150	1430	1670	1880	2220	2500

ILLUSTRATION 7F.(4). FACTORS BY WHICH WEIGHTED SAMPLE SIZES ARE MULTIPLIED TO OBTAIN APPROXIMATE UNWEIGHTED SAMPLE SIZES.

(Source: Indonesia Fertility Survey, 1976)

SUBCLASS	All Juwa-Bali	Type of Place		Province*				
		Urban	Roral	Jawa Barat	Jawa Tengah	Yogykarta	Jawa Timur	Bali
ALL	1.00	2.04	0.81	0.73	0.76	3.53	0.70	4.83
AGE								
Under 25	0.95	2.06	0.77	0.70	0.74	3.56	0.70	4.77
25-34	1.03	2.07	0.83	0.74	0.75	3.53	0.70	4.86
35-44	1.02	2.00	0.82	0.74	0.76	3.54	0.71	4.79
45-49	0.98	2.00	0.79	0.76	0.80	3.46	0.69	5.00
YEARS SINCE MARRIAGE								
Under 10	1.04	2.09	0.83	0.72	0.76	3.50	0.71	4.78
10-19	1.03	2.05	0.83	0.74	0.75	3.57	0.70	4.82
20-24	1.00	2.04	0.82	0.71	0.76	3.46	0.71	4.95
25 +	0.89	1.92	0.73	0.74	0.76	3.52	0.68	4.92
AGE AT MARRIAGE								
Under 15	0.78	1.99	0.70	0.69	0.71	3.64	0.65	5.17
15-19	0.90	2.02	0.82	0.71	0.73	3.66	0.69	4.82
20 +	1.43	2.12	1.14	0.87	0.89	3.46	0.77	4.85
LEVEL OF EDUCATION								
No schooling	0.93	—	—	0.65	0.71	3.63	0.65	4.81
Primary Incomplete	0.95	—	—	0.73	0.75	3.49	0.71	4.88
Primary Completed	1.14	—	—	0.79	0.92	3.37	0.81	5.00
Junior High +	1.73	—	—	1.16	1.27	3.20	1.08	4.90
HUSBAND'S OCCUPATION								
Prof., Admin, Clerical	1.45	—	—	0.94	1.06	3.46	0.87	4.93
Sales, Services	1.11	—	—	0.79	0.92	3.22	0.85	4.59
Manual	1.24	—	—	0.80	0.97	3.26	0.84	4.77
Farming	0.83	—	—	0.63	0.65	3.68	0.62	4.89

ILLUSTRATION 7G COMPONENTS OF VARIANCE

Total variance of a statistic may be decomposed into components in various ways such as according to sampling stages, steps in the estimation procedure, or some other features of the design (see Chapter 5). Information on variance components may not be of interest to the general user or substantive analyst, but it can be most useful for survey design work. Few surveys provide information on variance components; in part this is because of the complexity and difficulty involved in decomposing total variance into components. As a rare illustration, Tables 7G.(1) and (2) show some information from the US Current Population Survey (United States, 1968). The information has been presented concisely for a set of important variables: Table 7G.(1) showing components by sampling stage; and Table 7G.(2) by steps in the estimation procedure. The CPS sampling plan is essentially equivalent to dividing the entire US population into ultimate sampling units (clusters) each containing about 4 neighbouring housing units, and then selecting a clustered sample of the USUs. Table 7G.(1) shows that the main component consists of variance between USUs within PSUs, though for some variables the between-PSU component exceeds 10% of the total. The rather unusual and generally small 'between stratum' component refers to variance among strata totals within pairs of 'collapsed' strata into which the sample was divided for the purpose of variance computations.

The above analysis of variance applies to the 'final' estimates from the survey. These estimates were themselves produced in a number of estimation stages: (i) essentially unbiased estimates with adjustment for non-response; (ii) the first stage ratio estimation to reduce the contribution to variance arising from the selection of PSUs on the basis of past census information: (ii) the second stage ratio estimation which adjusts survey estimates in a number of age-sex-race groups to independently derived current estimates; and finally (iv) the composite estimates computed as weighted averages of current and previous months estimates. (For details of the procedure, see United States, 1968). Table 7G.(2) shows the ratio of the variance of the estimator obtained after each stage in the estimation procedure to the corresponding variance of the final composite estimator. This is the factor by which variance is reduced due to the subsequent stages in the estimation procedure. For instance for the variable 'Not in labor force, total', the variance of the unbiased estimator produced after step (i) is 4.07 times larger than the variance finally obtained for the composite estimator for this variable. The subsequent three stages reduce the variance by this factor. There is a very slight reduction due to first stage ratio estimation (step ii), from 4.07 to 4.01; but a major decrease, from 4.01 to 1.15 times the final variance after second stage ratio estimation (step iii).

ILLUSTRATION 7G.(1). COMPONENTS OF VARIANCE BY SAMPLING STAGE.

Components of Estimated Sampling Variance for CPS Composite Estimates of U S Level, Monthly Averages 1975

Population 16 years old and over	Average estimate of level[1] (x 10³)	Average standard error of level[2] (x 10³)	Distribution of variance[3] Within PSU (percent)	Between PSU (percent)	Between stratum (percent)
(1)	(2)	(3)	(4)	(5)	(6)
Not-in-labor-force total	*58 655*	*228 01*	*83 7*	*14 7*	*1 6*
Not white	7 239	85 02	95 0	5 1	–
Civilian labor force, total	*92 612*	*228 01*	*83 7*	*14 7*	*1 6*
Not white	10 529	85 02	95 0	5 1	–
Part time	8 197	112 68	96 4	2 6	1 0
Teenagers 16-19 years old	8.799	82 53	93 9	6 0	1
Employed in agriculture, total	*3 381*	*103 54*	*87 0*	*9 7*	*3 3*
Males	2 801	85 43	91 4	5 1	3 5
Not white	284	31 15	78 2	23 6	–1 8
Teenagers 16-19 years old	453	31 42	87 4	12 7	– 1
Employed in nonagriculture, total	*81,402*	*243 39*	*90 3*	*8 0*	*1 7*
Males	48 429	146 17	96 9	1 5	1 6
Not white	8 787	86 24	98 0	1 8	1
Farm residence	1 948	80 92	85 6	15 4	– 1 1
With a job not at work	5 007	107 41	97 3	2 9	– 2
Self-employed	5 626	96 85	88 7	11 0	3
Teenagers, 16-19 years old	6 593	80 27	96 2	3 9	– 1
Unemployed, total	*7,830*	*123 75*	*93 1*	*5 4*	*1 5*
Not white	1,459	55 17	95 5	5 1	– 5
Teenagers 16-19 years old	1,752	55 80	96 0	3 9	2
In SMSA's total	*103 355*	*446 32*	*93 8*	*4 4*	*1 7*
Central city	44,956	367 27	102 8	–4 6	1 8
Rural nonfarm residence	36 919	454 82	93 1	4 9	2 0
Rural farm residence	6 520	185 44	106 9	– 5 8	–1 2
Household heads	72 422	165 80	99 9	– 6	8

– Entry represents zero

[1] The arithmetic mean of the 12 monthly estimated levels for the year

[2] The square root of the arithmetic mean of the variances of the 12 monthly estimated levels for the year

[3] The percent of the estimated total variance attributed to the three stages of sampling (See p 96)

ILLUSTRATION 7G.(2). THE EFFECT OF ESTIMATION PROCEDURES ON VARIANCE.

Variance of Composite Estimates of Monthly U.S. Level and Variance Factors for Selected Estimators, Monthly Averages: 1975

Population 16 years old and over	Variance of composite estimate of level[1] (x 10^3)	Variance factor[2]			
		Unbiased estimator	Unbiased estimate with		
			First stage	Second stage	First and second stage
(1)	(2)	(3)	(4)	(5)	(6)
Not in labor force, total	*51 991*	*4 0698*	*4 0132*	*1 1543*	*1.1046*
Not-white	7 228	4 9031	4 0965	1 1022	1 0979
Civilian labor force total	*51 991*	*8 3244*	*8 2473*	*1 1543*	*1 1406*
Not-white	7.228	8 3173	7 3137	1 1022	1 0979
Part time	12 696	1 2010	1 2017	1 0859	1 0832
Teenagers 16-19 years old	6 811	3 2240	3 2128	1 0359	1 0306
Employed in agriculture, total	*10 720*	*1 4785*	*1 0603*	*1 5300*	*1 1076*
Males	7 299	1 5001	1 0460	1 5649	1 1060
Not-white	970	1 1583	1 0150	1 2478	1 1424
Teenagers, 16-19 years old	987	1 1842	1 0792	1 1633	1 0670
Employed in nonagriculture, total	*59 237*	*6 0689*	*6 0057*	*1 2151*	*1 1666*
Males	21 365	6 7699	6 6832	1 2189	1.1569
Not-white	7 437	5 6909	5 1624	1 2178	1 1762
Farm residence	6 547	1 2052	1.1749	1 2660	1 2306
With a job not at work	11 537	8752	8710	8496	8473
Self-employed	9 380	1 2662	1 2622	1 1835	1 1844
Teenagers, 16-19 years old	6 444	2 3790	2 3715	1 0841	1 0699
Unemployed, total	*15 313*	*1 1609*	*1 1578*	*1 0101*	*1 0007*
Not-white	3 044	1 2898	1 2074	9965	9758
Teenagers 16-19 years old	3 114	1 1254	1 1202	9669	9582
In SMSA's, total	*199 206*	*2 6815*	*2 6155*	*1 1950*	*1.1318*
Central city	134 885	1 5732	1 3271	1 4023	1 2137
Rural nonfarm residence	208 862	2 0687	1 6124	1 5997	1 1358
Rural farm residence	34 389	1 8227	1 2250	1 8465	1 2361
Household heads	27 490	6 6847	6 5833	1 2884	1 2706

[1] The arithmetic mean of the 12 variances of monthly level for the year

[2] The ratio of the variance of the indicated estimator to the variance of the composite estimate of level. Factors are based on monthly variances averaged for the year

REFERENCES

ALIAGA, A., and VERMA, V. (1991). An analysis of sampling errors in the Demographic and Health Surveys. DHS World Conference, Washington D.C.

BEAN, J.A. (1970). Estimation and sampling variance in the Health Interview Survey. National Centre for Health Statistics, Series 2, no 38. Washington D.C.: US Dept of Health, Education and Welfare.

CHINA(1986). In-depth Fertility Surveys in China: Principal Report. Beijing: State Statistical Bureau.

COCHRAN, W.G. (1973). Sampling Techniques (Third Edition). New York: Wiley.

CYPRUS (1990). Review of Sample Designs of Economic and Employment Surveys. Research Papers and Reports, Series II, no 14. Nicosia: Dept of Statistics and Research, Ministry of Finance.

DEMING, W.E. (1960). Sample Design in Business Research. New York: Wiley.

FULLER, W.A., SCHNELL, D., SULLIVAN, D., and KENNEDY, W.J. (1987). Survey variance computations on the personal computer. Bulletin of the International Statistical Institute, 52(3), 23-37.

PC CARP is available from Statistical Laboratory Iowa State University, Ames, Iowa, 50011, USA. Organisations in developing countries may contact International Statistical Programs Center, Bureau of the Census, Washington DC, 20233, USA.

GONZALEZ, M., OGUS, J., SHAPIRO, G., and TEPPING, B. (1975). Standards for discussion and presentation of errors in survey and census data. Journal of the American Statistical Association, 70 (no. 351, pt II).

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). Sample Survey Methods and Theory. New York: Wiley.

HANSEN, M.H., HURWITS, W.N., and BERSHAD, M. (1961). Measurement of errors in censuses and surveys. Bulletin of the International Statistical Institute, 38(2), 359-374.

INDIA (1990). Country Paper. ESCAP/PRC Seminar on Design and Evaluation of Household Sample Surveys, Beijing, China, May 22-28.

INDONESIA (1978). Indonesia Fertility Survey 1976: Principal Report. Jakarta: Central Bureau of Statistics.

INTERNATIONAL LABOUR OFFICE (1986). Statistical Sources and Methods Vol 3: Economically Active Population, Employment, Unemployment and Hours Worked (Household Surveys).

INTERNATIONAL LABOUR OFFICE (1990). Surveys on Economically Active Population, Employment, Unemployment and Underemployment: A Manual on Concepts and Methods.

INTERNATIONAL STATISTICAL INSTITUTE (1975). Manual on Sample Design. WFS Basic Documentation, 3.

KALTON, G. (1977). Practical methods of estimating survey sampling errors. Bulletin of the International Statistical Institute, 47(3) 495-514.

KALTON, G. (1979). Ultimate cluster sampling. Journal of the Royal Statistical Society A, 142(2), 210-222.

KALTON, G., and BLUNDEN, R. (1973). Sampling errors in the British General Household Survey. Bulletin of the International Statistical Insitute, 45(3), 83-97.

KAPLAN, B. and FRANCIS, I. (1979). Criteria for comparing programs for computing variances of estimators from complex samples. Proceedings of the 12th Annual Symposium on the Interface, Waterloo, Ontario, Canada.

KENDALL, M.G. and BUCKLAND, W.R. (1982). A Dictionary of Statistical Terms (4th Edition). London: Longman.

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

KISH, L. (1957). Confidence intervals for clustered samples. American Sociological Review, 22(2), 154-165.

KISH, L. (1965). Survey Sampling. New York: Wiley.

KISH, L. (1968). Standard errors for indexes from complex samples. Journal of the American Statistical Association,63, 512-529.

KISH, L. (1987). Statistical Design for Research. New York: Wiley.

KISH, L. (1989). Sampling Methods for Agricultural Surveys. Statistical Development Series, No 3. Rome: FAO.

KISH, L., and HESS, I. (1959). On variance of ratios and their differences in multistage samples. Journal of the American Statistical Association, 54, 416-446.

KISH, L., and FRANKEL, M. (1970). Balanced repeated replication for standard errors. Journal of the American Statistical Association, 65(331), 1071-1093.

KISH, L., and FRANKEL, M. (1974). Inference from complex samples. Journal of the Royal Statistical Society, B/36, 1-37.

KISH, L., GROVES, R., and KROTKI, K. (1976). Sampling Errors for Fertility Surveys. WFS Occasional Papers, No 17. The Hague: International Statistical Institute.

LAHIRI, D.B. (1957). Observations on the use of interpenetrating samples in India. Bulletin of the International Statistical Institute, 36(3), 144-152.

LITTLE, R.J.A. (1978). Generalised Linear Models for Cross-classified Data from the WFS. WFS Technical Bulletins, No 5. The Hague: International Statistical Institute.

MAHALANOBIS, P.C. (1944). On large-scale sample surveys. Phil. Trans. Royal Society, B, 231: 329-451.

McCARTHY, P.J. (1966). Replication: An approach to the analysis of data from complex surveys. Vital and Health Statistics, Series 2. No 14. US Dept of Health, Education and Welfare, Washington DC: Government Printing Office.

MURTHY, M.N., and ROY, A.S. (1975). Development of the sample design of the Indian National Sample Survey during its first twenty-five rounds. Sankhya, C(37), 1-42.

PLACKETT, R.L., and BURMAN, J.P. (1946). The design of optimum multifactorial experiments. Biometrika 33, 305-325.

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

RUST, K. (1987). Practical problems in sampling error estimation. Bulletin of the International Statistical Institute,, 52(3), 39-56.

RUSK, K., and KALTON, G. (1987). Strategies for collapsing strata for variance estimation. Journal of Official Statistics, 3(1), 69-81.

SCOTT, C., and HARPHAM, T. (1987). Sample design. In 'The World Fertility Survey: An Assessment', Cleland and Scott (eds). Oxford University Press.

SURVEY STATISTICIAN (1985-86 issues). Paris: International Association of Survey Statisticians.

TEPPING, B.J. (1968). Variance estimation in complex surveys. Proceedings of the Social Statistics Section, American Statistical Association, 11-18.

TUKEY, J.W. (1968). Discussion. Proceedings of the Social Statistics Section, American Statistical Association, p 32.

TURKEY (1980). Turkish Fertility Survey 1978: First Report. Ankara: Hacettepe Institute of Population Studies.

UNITED STATES (1968). The Current Population Survey: Design and Methodology. Technical Paper 40. US Dept of Commerce, Bureau of the Census.

UNITED STATES (1974). Standards for Discussion and Presentation of Errors in Data. US Bureau of the Census, Technical Paper No.32. Washington DC: Government Printing Office.

UNITED STATES (1978). An Error Profile: Employment as Measured by the Current Population Survey. Statistical Policy Working Paper No 3. Washington DC: Office of Federal Statistical Policy and Standards.

UNITED STATES (1982). Supplementary Report Provisional Estimates of Social, Economic and Housing Characteristics, PH C80-51-1. Washington DC: Government Printing Office.

UNITED NATIONS (1982). Survey Data Processing: A Review of Issues and Procedures. NHSCP Technical Study DP/UN/INT-81-041/1.

UNITED NATIONS (1982). Non-sampling Errors in Household Surveys: Sources, Assessment and Control. NHSCP Technical Study DP/UN/INT-81-041.2.

UNITED NATIONS (1989). Household Income and Expenditure Surveys. NHSCP Technical Study DP/UN/INT-88-XO1/6E.

VERMA, V. (1980). Sampling for national fertility surveys. Proceedings, World Fertility Conference, pp 383-436. The Hague: International Statistical Institute.

VERMA, V. (1982). The Estimation and Presentation of Sampling Errors. WFS Technical Bulletins, No 11. The Hague: International Statistical Institute.

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

VERMA, V. and PEARCE, M. (1986). CLUSTERS Version 3: Users' Manual. The Hague: International Statistical Institute.

CLUSTERS was originally supported and distributed by International Statistical Institute, 428 Prinses Beatrixlaan, 2270 AZ Voorburg, The Netherlands. It has been extensively used, and supplied to countries participating in the Demographic and Health Surveys, by Institute for Resource Development, Macro International Inc., 8850 Stanford Boulevard, Columbia, MD 21045, USA. Interested users may obtain CLUSTERS from the authors of the package by contacting them at 105 Park Road, Teddington, TW11 OAW, U.K.

WAKSBERG, J., HANSON, R., and BOUNPANE, P. (1973). Estimation and presentation of sampling errors for sample data from the 1970 US Census. Bulletin of the International Statistical Institute, 45(3), 66-82.

WOLTER, K. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

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

ZARCHOVICH, S.S. (1965). Sampling Methods and Censuses. Rome: FAO.

ZARCHOVICH, S.S. (1979) Stability of variance patterns. Journal of the Indian Society of Agricultural Statistics, 31(1), 23-48.