PN-AAM-301
ISN=23256
932054 4/62

WORLD FERTILITY SURVEY

TECHNICAL BULLETINS

DECEMBER 1982 NO. 11

The Estimation and Presentation of Sampling Errors

VIJAY VERMA

INTERNATIONAL STATISTICAL INSTITUTE
Permanent Office, Director: E. Lunenberg
428 Prinses Beatrixlaan, PO Box 950
2270 AZ Voorburg
Netherlands

WORLD FERTILITY SURVEY
Project Director:
Halvor Gille
35–37 Grosvenor Gardens
London SW1W 0BS, UK

The World Fertility Survey is an international research programme whose purpose is to assess the current state of human fertility throughout the world. This is being done principally through promoting and supporting nationally representative, internationally comparable, and scientifically designed and conducted sample surveys of fertility behaviour in as many countries as possible.

The WFS is being undertaken, with the collaboration of the United Nations, by the International Statistical Institute in co-operation with the International Union for the Scientific Study of Population. Financial support is provided principally by the United Nations Fund for Population Activities and the United States Agency for International Development. Substantial support is also provided by the UK Overseas Development Administration.

This paper is one of a series of Technical Bulletins designed to supplement the document 'Strategies for the Analysis of WFS Data', in the WFS *Basic Documentation* series. Technical Bulletins deal with specific methodological problems of analysis beyond the First Country Report.

For information on Country Reports, WFS publications, and WFS depository libraries, write to the Publications Office, International Statistical Institute, 428 Prinses Beatrixlaan, PO Box 950, 2270 AZ Voorburg, Netherlands. For information on the WFS generally, write to the Information Office, World Fertility Survey, International Statistical Institute, 35–37 Grosvenor Gardens, London SW1W 0BS, UK.

WORLD FERTILITY SURVEY

TECHNICAL BULLETINS

The Estimation and Presentation of Sampling Errors

VIJAY VERMA

United Nations Statistical Office
New York

DECEMBER 1982 NO. 11

Contents

ACKNOWLEDGEMENTS 5

1 INTRODUCTION: OBJECTIVES AND SCOPE 7

2 THE SIGNIFICANCE AND INTERPRETATION OF SAMPLING ERRORS 9

2.1 Introduction 9
2.2 Sampling Error and Other Survey Errors 9
2.3 Significance of Sampling Errors 11
2.4 Interpretation of Sampling Errors 13

3 PROCEDURES FOR ESTIMATION OF SAMPLING ERRORS 16

3.1 Introduction 16
3.2 A Practical Method of Computing Sampling Errors 16
3.3 Applications to Practical Designs 20
3.4 Sampling Errors for Complex Statistics 23
3.5 Variability of Variance Estimates 24
3.6 Confounding of Sampling and Response Variances in Computed Sampling Errors 25

4 PATTERNS OF VARIATION AND PORTABILITY 26

4.1 Objectives of Investigation 26
4.2 Portability 27
4.3 Modelling of Sampling Errors for Subclass Means 30
4.4 Sampling Errors for Subclass Differences 42
4.5 Design Factors for Complex Statistics 43
4.6 Extrapolation across Variables 44
4.7 Sampling Errors for Fertility Rates 45

5 PRESENTATION OF SAMPLING ERRORS IN SURVEY REPORTS 47

5.1 Modes of Presentation 47
5.2 For the General Reader 48
5.3 For the Substantive Analyst 50
5.4 For the Sampling Statistician 52

REFERENCES 54

APPENDIX A – A BRIEF DESCRIPTION OF THE CLUSTERS PACKAGE 56

TABLES

1 Deft values for small selected subclasses and subclass differences, compared to estimated increase in standard error due to departure from self-weighting 32

2 Comparison of (a) computed and (b) predicted subclass standard errors for selected variables from the Turkish Fertility Survey 34

3 Comparison between computed and predicted subclass standard errors for the estimated mean number of children ever born, by age group of women 35

4 Standard errors for subclass aged 15–19: comparison between computed and predicted values for proportions 36

5 Pattern of results for subclasses and subclass differences, by country, variable group and subclass group, averaged over selected variables and subclasses 38

6 Fitting of relation (4.18) to each of the eight domains by type of place of residence in the Turkish Fertility Survey 39

7 Comparison of (a) computed and (b) predicted standard errors for geographic domains in the Turkish Fertility Survey 40

8 Illustration of extrapolation procedure for estimating sampling errors for subclasses 41

9 Approximate value of standard error, by variable and subclass size (n_s), Indonesia Fertility Survey, 1976 51

10 For standard error (se_d) of the difference between two subclasses of size n_1 and n_2, the appropriate sample base (n_d) to be used in table 9 52

11 Factor by which weighted frequencies should be multiplied to obtain the corresponding unweighted sample size for various subclasses of the sample, by province and type of place of residence 53

Acknowledgements

This bulletin draws on the collective work of several colleagues at the World Fertility Survey. Among them, special thanks are due to Alan Sunter, who collaborated on an earlier draft of the document, and to Chris Scott, Colm O'Muircheartaigh and John McDonald for review and helpful suggestions. The opinions expressed in this work are my own, and not necessarily those of the United Nations.

1 Introduction: Objectives and Scope

This is one of a series of Technical Bulletins issued by the World Fertility Survey with the objective of illustrating applications of statistical methodology to various aspects of the analysis of sample survey data, and in particular of WFS data.

The present bulletin is concerned with the *estimation* and *interpretation* of sampling errors of survey estimates. Consideration is also given to the question of *presentation* of sampling errors in survey reports in a way which facilitates their proper use by researchers in the interpretation of substantive results, as well as in sample design and evaluation.

It has been a long-standing practice of the WFS to encourage and assist participating countries in publishing detailed sampling error estimates along with substantive results of the survey. Considerable effort has been made in this direction. For example, the WFS 'Data Processing Guidelines' (1980) contain recommendations on how to code the sample structure to ensure that sampling errors can be computed, and provide detailed specification of survey variables and sample subclasses for which computation of sampling errors is recommended. The WFS has developed (and distributed at a nominal charge) a package program, CLUSTERS, suitable for routine and large-scale computation of sampling errors for descriptive statistics from complex samples (Verma and Pearce 1978). Comparative analysis of sampling errors from a number of fertility surveys has also been undertaken (Kish, Groves and Krotki 1976; Verma, Scott and O'Muircheartaigh 1980). Consequently, practically all First Country Reports of WFS surveys include detailed sampling error estimates, and many provide excellent examples of procedures for estimation, presentation and interpretation of sampling errors.

Drawing on this work, the present bulletin aims at providing more systematic and detailed guidelines on computation, presentation, interpretation and use of sampling errors. Section 2 defines sampling error, placing it in the context of the total survey error, and considers why it is useful to compute sampling errors. It also provides a simple exposition of the interpretation and use of sampling errors, with illustrations. This section is directed specifically to the general user of survey results who, in reaching conclusions from the survey, must take into account the quality of the data and the associated margins of uncertainty, including those due to sampling variability.

The next three sections are directed specifically at the statistician and subject matter specialist responsible for the production of survey reports; these sections should also be useful in enhancing the understanding of the general reader of survey reports. Section 3 describes practical methods of computing sampling errors. The emphasis is on general and simple procedures which provide reasonably good approximation in diverse situations and hence are suitable for routine and extensive computations. The context here, as elsewhere, is that of a large-scale, single-round survey, with a probability sample and complex design, aimed at providing a variety of descriptive statistics of the type encountered, for example, in WFS First Country Reports (WFS 1977). Section 4 explores patterns of variation in sampling error results, across sample subgroups and across substantive variables, in the light of theoretical and empirical considerations. The objective is to illustrate how information on sampling errors may be summarized, and also extrapolated to subclasses, variables and samples other than those for which actual computations are performed. Section 5 provides guidelines on presentation of sampling error results for different types of users: the general reader and user of survey results, the subject matter specialist engaged in primary and secondary analysis of the survey data, and the sampling statistician interested in evaluating the design used for guidance in designing other, future samples.

Previous Page Blank

Finally, a brief outline of the package program CLUSTERS is provided in appendix A. The availability of this package is one of the factors which have made it possible for most WFS First Country Reports to include information on sampling errors, and organizations undertaking other surveys may profitably utilize the package for the same purpose.

The bulletin has a dual objective: to provide a step-by-step guide on how to compute and utilize sampling errors for diverse statistics, and at the same time to enhance the reader's appreciation of the nature and significance of errors or uncertainties inherent in the sampling process. The emphasis of the discussion is on its possible relevance to practical survey situations rather than on theoretical refinements. Specialists may regard some of the material included as common knowledge available in many excellent textbooks; however the document is aimed at a wider body of survey practitioners and users of survey data who have to take many decisions without being themselves expert sampling statisticians.

2 The Significance and Interpretation of Sampling Errors

2.1 INTRODUCTION

It is widely recognized as good practice for survey reports to include detailed information on sampling variability of the survey estimates. Cross-tabulations of the data from large-scale multi-purpose surveys generally involve numerous estimates over diverse subgroups, each with its associated sampling error. This section is concerned with the basic question: how the general reader and practical user of the survey results can utilize information on sampling errors in interpreting the substantive survey results and in drawing inferences from the survey.

Errors in surveys arise from numerous sources and sampling error is just one component of error. To appreciate its significance, it is important to place it in the context of the *total error*. Section 2.2 outlines a typology of survey errors and defines exactly what component of the total error is referred to as the 'sampling error'. In this context, the remainder of section 2 deals with the practical question: what can the user of survey results do with the information on sampling errors? Section 2.3 discusses the significance of sampling error in survey design and the interpretation of survey results, and sections 2.4 and 2.5 describe and illustrate how information on sampling error may be interpreted as margins of uncertainty in the estimates obtained from a sample survey.

2.2 SAMPLING ERROR AND OTHER SURVEY ERRORS

The objective of a sample survey is to make estimates or inferences of general applicability for a population, on the basis of observations made on a limited number (sample) of units of the population. We may define 'error' as the difference between the (usually unknown) actual population value and the value estimated from the observed sample, and we distinguish two broad categories of error (Verma 1981a). First, errors arise from the fact that what is observed or measured departs from what it is intended to measure in the survey. Such errors of measurement centre on the substantive content of the survey: definition of the survey objectives, their operationalization into a coherent and consistent set of questions, the interpretation and communication of these to the respondent, the respondent's ability and willingness to provide the information sought, the quality of recording, editing and coding the responses, etc.

Secondly, errors arise from the process of extrapolation of the results from the observed units to the entire study population. These errors centre on the process of sample design and selection, and will be present even if there are no errors of measurement involved in the units actually enumerated. It is important to define this second group of errors since sampling error, which is a component of it, is sometimes confused with the group as a whole.

To draw inferences from a sample survey about the population under study in an objective manner, it is necessary to have a probability sample; that is, it is necessary for the relative chance of being selected into the sample to be known and non-zero for each unit in the population.

Even when the sample is designed to be a probability sample, the above conditions may be violated in practice due to defects in sample implementation. Operationally, a sample is selected from a frame which explicitly or implicitly provides a list, and the sample design provides specific rules and procedures for sample selection from this list.

However, it may happen that not all units in the study population are included in the operational sampling frame (non-coverage); or some units may be duplicated in the frame, increasing their chance of selection into the sample; or there may be insufficient information for unique definition and identification of units in the field, or ambiguities in the correspondence between units of sampling and units of enumeration; or sample selection may not be executed as intended or designed; or information may not be obtained from all the units specified as belonging to the sample (non-response). As a consequence of factors such as these, the sample structure can be distorted by some units being entirely excluded or selected with probabilities different from those required by the sample design. These errors of sample implementation (as well as any deliberate departures from probability sampling) can result in known or unknown biases in estimates obtained from the survey.

Errors of measurement and errors of sample implementation are collectively termed *non-sampling errors*. As sketched above, these arise from a wide variety of sources and affect the survey results in different and complex ways (for a detailed discussion, see United Nations 1982). As distinct from these, the *sampling error* is inherent in the process of statistical estimation of population parameters from results obtained on a probability sample of the population. A sample design specifies rules by which units from the population are to be selected for enumeration and rules for the estimation of population parameters; even in the absence of measurement and implementation errors, repeated applications of the same design would result in different estimates depending on the actual units which happened to be selected. The sampling error of an estimator is a measure of its variability under the theoretically possible repetitions of the survey in the absence of non-sampling errors.

In general, the effect of any particular source of error (sampling or non-sampling) on aggregate survey results can be decomposed into two components: *variable error* and *bias*. This distinction is based on the possibility, in principle, of repetition of the survey under the same conditions. Some of the conditions under which the survey is taken may be considered *essential* to its design and operation: for example, the general social conditions, characteristics of the population enumerated, quality of the sampling frame available, nature and complexity of the information sought, or the type of survey staff and other facilities available. In addition to the essential conditions, survey results are also influenced by transient or chance factors, such as the particular units selected into the sample, the particular field and office staff used, or the conditions under which a particular interview is conducted. One can conceive of the survey being repeated under the given essential conditions; if this were done, different repetitions would still give different results due to the varying impact of chance factors. The variable component of the error measures the variability between different estimates made from such hypothetical repetitions of the survey. The average of all possible repetitions is the *expected value* under the given essential conditions. The difference between this expected or average value and the desired population value is the *bias* (Hansen, Hurwitz and Bershad 1961). More simply, but approximately, bias is the constant component of error which has the same effect on any repetition of the survey (Kish 1965: 517).

Variable error measures variation between estimates from different repetitions under the same essential conditions. The possible repetitions may be considered to be made up of two 'layers': repeated observations over a fixed sample of units; and repetition of the survey over different samples. Measurement variance, or response variance, is a measure of the variability of repeated measurements over the *same* sample of units; the variability of the average or expected value of these measurements over *different* samples is the sampling variance.

In short, sampling variance is intended to measure the variability between estimates

from different samples, to the exclusion of variable errors and biases resulting from the processes of measurement and sample implementation. (However, as explained in section 3, *estimates* of sampling variance in practice often include some contribution from other sources of variability, such as response variance.)

Finally we may note that the sampling 'mean square error' is composed of the sampling variance plus the square of the *sampling or estimation bias.* The latter bias is defined as the difference between the expected value of an estimator and the population parameter being estimated, in the absence of measurement errors and sample implementation errors. This bias is purely a statistical property of the estimator used, and with reasonable sample size and design can usually be avoided through the use of proper estimation procedures, such as by appropriately weighting sample values to compensate for differences in probabilities of selection. It is in this sense that most estimates considered in WFS First Country Reports are 'unbiased' (see, for example, Central Bureau of Statistics, Indonesia, 1978: 131). However, in certain situations the estimation bias may become serious, as in the case of ratio estimates from clustered samples with very unequal cluster sizes, as described in section 3.

2.3 SIGNIFICANCE OF SAMPLING ERRORS

In interpreting information on sampling errors, the reader has to bear in mind that they represent only one component of the total survey error. For estimates based on a relatively small sample size, this component may be the dominant one; however, in other situations, non-sampling errors, particularly systematic biases, may be much more important. In surveys with considerable rates of non-response, refusal, response error, listing error, etc, it is not always easy to decide how much attention should be given to sampling errors. Some survey statisticians give the impression of being exclusively concerned with the sampling errors, which are often easily computed, and ignoring the possibly more significant, but often unknown, non-sampling errors. Such an orientation is obviously not defensible. On the other hand, it is equally meaningless to declare that *in general* non-sampling error is predominantly more important than sampling error since the latter increases progressively as the size of the population subgroup under consideration diminishes. Thus in a small enough subgroup, the sampling error is almost certain to outweigh the non-sampling error.

It is important to appreciate the significance of information on sampling errors. As an experienced statistical organization has noted (Yugoslavia Federal Statistical Office 1978):

> Only after years of experience with a variety of surveys have we come to a firm opinion that *sampling errors have an orientational value.* Data from a sample survey might be, at least in principle, anything between 'excellent' and 'useless'. An inspection of the magnitude of sampling errors for various characteristics at the level of the country as a whole as well as its subdivisions is the first step in passing judgement about the place of the survey between these extremes. Therefore, in order to establish the basis for the evaluation process of the sample survey data, information about the magnitude of sampling errors should be considered as an *indispensible* part of each sample survey report.
>
> Needless to say, a knowledge of sampling errors is no more than a part of the information needed for the evaluation process.
>
> At its subsequent stages this process has to go into biases, the enumerator effect, a general study of the precautionary measures taken in the field, etc. However, all these additional steps have a very limited value unless they are combined with information about sampling errors [italics in the original.]

To the above consideration of orientation it must be added that information on the magnitude of sampling errors is essential in deciding the degree of detail with which the survey data may be meaningfully classified. In a fertility survey, for example,

interpretation of survey results generally requires very detailed classification of the data by demographic characteristics such as age, marriage duration or parity, and by the socio-economic background characteristics of the respondent. Even for a sample of a few thousand respondents, the sample categories being compared and contrasted can rapidly become very small. Roughly speaking, while the magnitude of non-sampling bias in a category is more or less independent of its sample size, its sampling variance tends to increase with decreasing sample size. Consequently the sampling error can be the predominant, or at least a significant, part of the total error for many small categories and comparisons of substantive interest.

Sampling error information is also essential for sample design and evaluation. Of course sample design is severely constrained by a variety of practical considerations such as the availability of sampling frames, fieldwork arrangements, the survey timetable, requirements of supervision and control, and, above all, the survey budget. (See WFS, 'Manual on Sample Design' (1975) for a useful exposition of the factors involved in the choice of sample design for fertility and similar surveys.) 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 was 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. This is illustrated by an evaluation of sampling errors from WFS surveys which concluded that 'perhaps there has been too strict an adherence to proportional allocation between domains . . . In some of the more heterogeneous, large countries . . . greater emphasis should have been given to survey estimates at the subnational level'. Further, 'it is possible that the WFS has erred in the direction of over-scattered sample designs. Certainly in some countries the use of more heavily clustered samples would have been more economical' (Verma 1981a).

2.4 INTERPRETATION OF SAMPLING ERRORS

In section 2.2 the concept of sampling error was defined in terms of variability between estimates from different samples. In this section measures of sampling error such as variance, standard error and confidence intervals will be defined and interpreted more precisely.

Suppose that $\bar{y}_s$ is a certain quantity such as a mean estimated from a particular sample s (it is assumed throughout for convenience that no measurement or other non-sampling errors are present). Different samples would give different values of $\bar{y}_s$; the distribution of $\bar{y}_s$ from all possible samples (with given design and selection probabilities) is called its *sampling distribution*. The sampling distribution of $\bar{y}_s$ is the theoretical distribution of the estimate over all possible samples, each sample weighted by its probability of occurrence, P_s, depending upon the sample design applied to a fixed population of characteristics. The expected value of the estimate is the mean of the sampling distribution:

$$E(\bar{y}) = \sum_s P_s \cdot \bar{y}_s \tag{2.1}$$

where $\sum_s$ is the sum over all possible samples.

The *variance* of $\bar{y}$ is measured by the square of the difference between a sample estimate $\bar{y}_s$ and its expected value over all possible samples, $E(\bar{y})$, averaged over all possible samples, ie

$$\mathrm{Var}(\bar{y}) = \sum_s P_s \cdot [\bar{y}_s - E(\bar{y})]^2 \tag{2.2}$$

The *standard error*, $\sigma_{\bar{y}}$, is the square root of the variance. The sampling distribution represents the random fluctuations of $\bar{y}_s$ due to the specific sample design, and this variability is measured by the standard error.

In a practical situation, results from only one sample are available. However, inherent in a properly designed probability sample is the ability to provide estimates of the sampling error from the results of the one sample that is available. This is because the observed variability between units within a sample can provide an estimate of variability between different samples. It should be appreciated that the estimated standard error, say $se(\bar{y})$, from a particular sample does not measure the actual deviation $\bar{y}_s - E(\bar{y})$ of that particular sample mean from the expected value; rather it is an estimate of a parameter $\sigma_{\bar{y}}$ of the sampling distribution of this deviation. In fact it is not necessary for $\bar{y}_s$ and $se(\bar{y})$ to be estimates from the same data.

Inferences from sample surveys are made in terms of probability intervals, usually *confidence intervals*. These intervals are defined on the basis of the sampling distribution of $\bar{y}_s$ (ie the distribution of the estimates $\bar{y}_s$ obtained from all possible samples with given design and selection probabilities). In many practical situations this distribution is approximately normal. For values of $\bar{y}_s$ distributed normally around their mean $E(\bar{y})$ with standard deviation $\sigma_{\bar{y}}$, the probability P of $\bar{y}_s$ being in the interval $\pm$ t times the standard error on either side of the expected value, ie the interval $E(\bar{y}) \pm t \cdot \sigma_{\bar{y}}$, is given by the following table:

t	0.67	1.00	1.64	1.96	2.58	3.00	3.29
P	0.50	0.68	0.90	0.95	0.99	0.997	0.999

For example, 68 per cent of sample estimates $\bar{y}_s$ are expected to lie within the range $E(\bar{y}) \pm \sigma_{\bar{y}}$; similarly 95 per cent lie within the range $E(\bar{y}) \pm 2 \cdot \sigma_{\bar{y}}$ approximately.

In practice we interpret such an interval as follows. The estimated confidence interval

$$\bar{y}_s \pm t \cdot se(\bar{y}) \tag{2.3}$$

contains the expected value $E(\bar{y})$ with probability (or confidence) P, where, for the assumed normal distribution, the relationship between t and P is given by the above table.

As noted above, in many practical situations the sampling distribution of $\bar{y}_s$ is approximately normal. Just how good that approximation is depends on the underlying distribution of the characteristics of the population and on the sample design. The approximation improves with increasing sample size and, for most samples encountered in practical survey research, the assumption of normality leads to errors that are small compared to other sources of error. Note that the fact of this approach to normality of the sampling distribution of large samples does not depend on the normality of the distribution of elements in the population. The distribution of many characteristics in the population is, in fact, far from normal. For example, the number of children ever born to married women in a cross-sectional survey may be highly skewed to the left; however the *means* estimated from different samples of reasonable size are likely to be approximately normally distributed around the expected value.

For clustered samples based on a small number of clusters (say less than about 30), it is more appropriate to use the Student t-distribution, with 'degrees of freedom' (df) equal to

$$\sum_h (a_h - 1) = \sum_h a_h - H,$$

where a_h is the number of clusters in stratum h and $\sum_h$ the sum over all H strata, ie df is equal to the total number of clusters less the number of strata. For example, for a sample of 20 clusters with 2 selections per stratum from 10 strata,

$$df = 20 - 10 = 10$$

and the confidence intervals are

t(df = 10)	0.70	1.05	1.81	2.23	3.17	3.96	4.59
P	0.50	0.68	0.90	0.95	0.99	0.997	0.999

Compared with the normal distribution (which corresponds to infinite df), the above distribution gives a broader interval (ie a larger value of t in equation 2.3 for the same level of confidence), for example the 95 per cent of confidence interval is $\bar{y}_s \pm 2.23$ $SE(\bar{y})$, compared with $\bar{y}_s \pm 1.96\,SE(\bar{y})$ for the normal distribution. This difference is larger at higher values of t.

Some statisticians prefer to publish information on sampling errors in terms of the standard error, to permit the reader to construct intervals and make inferences according to his needs. However, for the general reader it is more useful to know the range within which the 'true' value of interest can be expected to lie with a certain level of confidence. While different levels of confidence may be chosen for different purposes, a common and convenient criterion (followed, for example, in most WFS First Country Reports) asserts that the population value to be estimated from the sample lies within a range

twice the standard error on either side of the sample value. This can be asserted with a high (95 per cent) level of confidence, ie one can say that the chances are only one in twenty that the true value is outside this range. For example, if the observed sample mean for a variable is 3.5 and the estimated standard error is 0.2, then for practical purposes, apart from non-sampling errors and other biases, the true population value of interest lies in the range

$$3.5 \pm 2\,(0.2) = 3.1 \text{ to } 3.9$$

with 95 per cent confidence.

The question of whether or not two subgroups of the sample differ significantly in a particular characteristic can be dealt with in a similar way. One first obtains the observed difference, found in the sample; one then estimates the standard error of the difference and this leads to the 95 per cent confidence limits for the difference. If a difference of zero would fall outside these limits, then it can be said that the hypothesis of no difference is rejected with 95 per cent confidence, or, in other words, that the groups differ significantly.

As an example, suppose that two group means are being compared:

Group 1 observed mean = 3.5

Group 2 observed mean = 3.0

Observed difference: 3.5 − 3.0 = 0.5

Suppose the standard error of the difference is estimated at 0.15. Then one can assert (with 95 per cent confidence) that the true difference is in the range

$$0.5 \pm 2\,(0.15) = 0.2 \text{ to } 0.8,$$

and the observed difference is said to be 'statistically significant', because we have more than 95 per cent confidence that it is not zero.

For a lucid discussion on the use and misuse of tests of significance in social research, see Kish (1959).

3 Procedures for Estimation of Sampling Errors

3.1 INTRODUCTION

Practical methods of computing sampling errors need to be *general* enough to cover the complexities and variations which frequently arise in sample designs for large-scale surveys. The estimation procedure must take into account the sample structure, in particular its clustering and stratification. At the same time the procedures should be *simple* enough to provide easy computational formulae so as to permit, economically, the detailed computations required for the numerous estimates produced by the survey.

In a multi-stage design, each stage of selection makes a contribution to the total sampling error. The contribution of the first stage results from the fact that only some of the first-stage, or primary, sampling units (PSUs) in the population are taken into the sample. The contribution of the second stage results from the fact that only some of the second-stage units (SSUs) from within the selected PSUs appear in the sample, and so on. For sample design and for the evaluation of sample designs, decomposition of the total sampling error into its components according to sampling stages (along with information about costs, etc) may be required. However, as a guide for orientation towards and interpretation of survey results, the user requires information on the overall magnitude of the sampling error, without requiring its decomposition into contributions of individual stages. This section describes practical methods for estimating the overall sampling error for complex samples.

The essential procedure for estimating sampling errors for complex samples is illustrated by the use of simple replicates or interpenetrating samples first introduced by Mahalanobis (1944). If the sample is divided into independent subsamples or replications, each of exactly the same design, then each of the subsamples yields a valid estimate of the population parameter of interest. A measure of variability among the replications provides an estimation of the sampling variance. For example, if the sample is divided into c independent replications, and y_i is the estimate of a sample total from replication i, then the estimated variance of the averaged estimate of the total

$$\bar{y} = \sum_i y_i/c$$

is

$$\mathrm{var}(\bar{y}) = \frac{1}{c(c-1)} \sum_i (y_i - \bar{y})^2 \qquad (3.1)$$

The use of interpenetrating samples provides an easy and convenient method of estimating variance, irrespective of the complexity of the design within any replication. Essentially the same approach is used for estimating sampling errors for other complex sample designs. For example, each independently selected primary sampling unit (PSU) from a stratum can provide a valid estimation for the stratum, so that a measure of variability among the PSUs within strata provides an estimate of the sampling variance. The procedure is described in detail in the remainder of this section.

3.2 A PRACTICAL METHOD OF COMPUTING SAMPLING ERRORS

Under certain assumptions, usually not too restrictive in practical situations, sampling errors for a variety of statistics, such as proportions, means, ratios and their differences,

over the total sample as well as over diverse subclasses, can be obtained on the basis of values totalled at the level of primary sampling unit, ie on the basis of PSU totals. The sample design and selection within individual PSUs may be complex and differ from one PSU to another, without affecting the form of the variance estimation to be described below. *The method takes into account the components of variance from all, including the second and subsequent, sampling stages, even though no explicit reference appears in the computational formulae to any stage beyond the first.* Essentially this is because the variance contributed by the later stages is reflected in the observed variation among first-stage units. A review of practical methods of computing sampling errors is provided by Kalton (1977).

The basic assumptions required are (a) that two or more PSUs are drawn from each stratum, and (b) that these selections are drawn independently of one another, with random choice and with replacement. These conditions are seldom satisfied exactly in practical sample designs; however, as described later (section 3.3), they are reasonably well approximated in many situations.

Suppose that the total sampling frame is divided into a number of strata, that PSUs are selected independently with replacement from each stratum, and that a subsample is selected in each selected PSU, in whatever way, so as to give a final sample of ultimate units.

Let y_{hij} be the value for ultimate unit j in PSU i from stratum h and let w_{hij} be the weight associated with the unit (introduced to compensate for unequal probabilities of selection, differential non-response, etc). Then

$$y_{hi} = \sum_j y_{hij} \cdot w_{hij} \tag{3.2}$$

is the appropriately weighted estimated total for the sample selected from PSU i, scaled in such a way that

$$y_h = \sum_i y_{hi} \quad \text{and} \quad y = \sum y_h \tag{3.3}$$

are the estimated stratum total and sample total, respectively. The variance of the totals is estimated as

$$\left.\begin{aligned} \operatorname{var}(y_h) &= \frac{a_h}{a_h - 1} \cdot \sum_i (y_{hi} - y_h/a_h)^2 = \frac{a_h}{a_h - 1}\left(\sum_i y_{hi}^2 - \frac{y_h^2}{a_h}\right) \\ &\text{and} \\ \operatorname{var}(y) &= \sum_h \operatorname{var}(y_h) \end{aligned}\right\} \tag{3.4}$$

where a_h is the number of PSUs selected from stratum h. The reader may note that the above is analogous to the formula for estimating the variance of a sample total for a stratified random sample.

Variance of Ratios

Generally, sample surveys are used to estimate ratios (rather than totals), of the form

$$r = \frac{\Sigma w_{hij} \cdot y_{hij}}{\Sigma w_{hij} \cdot x_{hij}} = \frac{\Sigma y_{hi}}{\Sigma x_{hi}} = \frac{\Sigma y_h}{\Sigma x_h} = \frac{y}{x} \tag{3.5}$$

For example, in a sample of women y_{hij} might be the number of living children to woman j (in PSU i, stratum h), and x_{hij} her total number of children ever born; then the ratio r estimates the proportion of surviving children.

Ordinary means and proportions are just special cases of ratios. In a mean, the denominator is simply a count variable, ie, $x_{hij} = 1$ for all ultimate units concerned. For a proportion, in addition the numerator is a dichotomy, with $y_{hij} = 1$ or 0 depending on whether or not the unit possesses the characteristic whose proportion is being estimated. Since ratios and their differences are commonly encountered estimates from sample surveys, it will be useful to list formulae for the estimation of their variance. For a detailed treatment with numerous numerical illustrations, see Kish and Hess (1959).

The variance of a ratio r, to a certain degree of approximation (see below), is

$$\text{var}(r) = \frac{1}{x^2}\left[\text{var}(y) + r^2 \cdot \text{var}(x) - 2r \cdot \text{cov}(x, y)\right] \tag{3.6}$$

where var (y) is given by equation 3.4, with a similar expression for var (x), and the covariance is

$$\text{cov}(x, y) = \sum_h \frac{a_h}{a_h - 1}\left(\sum_i y_{hi} \cdot x_{hi} - \frac{y_h \cdot x_h}{a_h}\right) \tag{3.7}$$

The same expressions can be used to compute the variance of r over a subclass of the sample: any units not belonging to the subclass are simply ignored.

Frequently the denominator is a count variable, so that x is the total sample size. The terms var (x) and cov (x, y) appear in equation 3.6 because of variation in cluster sizes.[1] Apart from its contribution to var (r), the variation in cluster size also affects the statistical bias in the ratio estimator. It has been shown that the bias relative to the standard error is less than the coefficient of variation of x, ie

$$\text{cv}(x) = \sqrt{\text{var}(x)}/x = \text{se}(x)/x$$

and decreases with the number of sampling units. In a well-designed sample, with little variation in cluster size, the bias is usually negligible. However, for subclasses which dic not form explicit strata for sample selection, the effective cluster size may vary greatly, as may be the case with very small or ill-distributed subclasses of the sample. For a stratified clustered sample, cv(x) is estimated by

$$\text{cv}^2(x) = \frac{1}{x^2}\left[\sum_h \frac{a_h}{a_h - 1}\left(\sum_i x_{hi}^2 - \frac{x_h^2}{a_h}\right)\right]$$

Variance of the Difference between Two Ratios

Turning now to the difference of two ratios

$$r - r' = \frac{y}{x} - \frac{y'}{x'} \tag{3.8}$$

its variance is given by

$$\text{var}(r - r') = \text{var}(r) + \text{var}(r') - 2\,\text{cov}(r, r') \tag{3.9}$$

where the variance terms are given by equation 3.6, and

$$\text{cov}(r,r') = \frac{1}{x \cdot x'}\left[\text{cov}(y,y') + r \cdot r'\,\text{cov}(x,x') - r \cdot \text{cov}(y',x) - r' \cdot \text{cov}(y,x')\right] \tag{3.10}$$

with cov(x, y), etc defined as in the form 3.7.

[1] The term 'cluster size' is used here to refer to the number of ultimate sampling units selected in the PSU.

A difference of ratios can arise in a number of ways. One may compare, for example, two different characteristics (y and y') over the same sample ($x = x'$). The most common situation, however, is the comparison of the same characteristic (same variate in the numerator) between two subclasses of the sample defined in terms of different categories of the same characteristic in the denominator. The classes are usually mutually exclusive but not necessarily exhaustive, for example, comparison of mean age at marriage among two age groups of women, eg 25–34 and 35–44, or comparison of the proportion of children dead between urban and rural women. The same computational formulae apply in different situations. The covariance term in equation 3.9 arises because the same ultimate units appear in both ratios or because the units come from the same sample clusters. The statistical bias in $(r - r')$ can become large in relation to $\sqrt{\text{var}}\,(r - r)$ if the covariance term is large and positive, or if the biases in r and r' are very different.

Other Functions of Ratios

By introducing the transformation

$$z_{hij} = \frac{1}{x}(y_{hij} - r \cdot x_{hij}) \tag{3.11}$$

equation 3.6 can be expressed more simply as

$$\text{var}\,(r) = \sum_h \left[\frac{a_h}{a_h - 1} \left(\sum_i z_{hi}^2 - \frac{z_h^2}{a_h} \right) \right] \tag{3.12}$$

where

$$z_{hi} = \sum_j w_{hij} \cdot z_{hij} = y_{hi} - r \cdot x_{hi} \quad \text{and} \quad z_h = \sum z_{hi} = y_h - r \cdot x_h,$$

and

$$z = \sum_h z_h = y - r \cdot x = 0 \quad \text{by definition.}$$

In fact equation 3.12 holds for the ratio of two ratios, the product of two ratios, and for the difference or any linear combination of ratios, with appropriately defined variable z_{hij}.

For a difference of two ratios $r'' = (r - r')$, equation 3.12 gives var $(r - r')$ with

$$\begin{aligned} z''_{hij} &= \frac{1}{x}(y_{hij} - r \cdot x_{hij}) - \frac{1}{x'}(y'_{hij} - r' \cdot x'_{hij}) \\ &= (z_{hij} - z'_{hij}) \end{aligned} \tag{3.13}$$

Similarly for the sum of two ratios $r'' = (r + r')$, we have

$$\begin{aligned} z''_{hij} &= \frac{1}{x}(y_{hij} - rx_{hij}) + \frac{1}{x'}(y'_{hij} - r'x'_{hij}) \\ &= (z_{hij} + z'_{hij}) \end{aligned} \tag{3.14}$$

The generalization of the above to any linear combination of ratios is straightforward.

For the ratio of two ratios (double ratio), $r'' = r/r'$, we have

$$z''_{hij} = \frac{1}{r'}(z_{hij} - r'' \cdot z'_{hij}) \tag{3.15}$$

and for the product of two ratios, $r'' = r \cdot r'$

$$z''_{hij} = (r' \cdot z_{hij} + r \cdot z'_{hij}) \tag{3.16}$$

An example of a double ratio is a 'relative fertility rate'. A fertility rate is itself a ratio, the numerator consisting of the number of births to a specified category of women during a specified interval of time, and the denominator being the number of person-years of exposure to childbearing for those women during the same interval. One might be interested in the distribution of fertility rates for different ages, or different categories of some kind, rather than their absolute values. The distribution is given by the 'relative fertility rate', ie the fertility rate of any category (such as an age group) divided by the overall rate. This is a double ratio.

An example of the product of two ratios is provided by age-specific fertility rates computed from combined data coming from a household survey and a fertility survey of a sample of ever-married women, as described in section 4.7 below.

3.3 APPLICATIONS TO PRACTICAL DESIGNS

The simplified variance estimation procedures described above are based on the assumption that two or more PSUs are selected independently and with replacement from each stratum. Frequently, actual designs do not satisfy these conditions exactly, and certain approximations are involved in fitting the model to them.

(1) Sampling without Replacement

In most practical situations it is preferable to select PSUs without replacement (ie without allowing any unit to appear more than once in the sample), since the resulting variance may be somewhat smaller than that obtained using sampling with replacement. Consequently, the procedure described above for with-replacement sampling would tend to overestimate the variance of a without-replacement sample. However, taking this feature of the design into account would require additional computation of other components of the variance and the use of complicated formulae.

Fortunately in most situations the overestimation involved is entirely trivial; when that is not the case, a better approximation is often provided by introducing the 'finite population correction' $(1 - f_h)$ into equation 3.4, ie

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

where f_h is the overall sampling fraction in stratum h.

(2) Systematic Sampling

Systematic sampling[2] serves as a practical and convenient method of selecting units from an ordered list; often the ordering for selecting PSUs is geographical, as is the case in most WFS samples. The combined effect of ordering and selecting by applying a constant interval to the list can be seen as introducing 'implicit' stratification, with one random selection from each implicit stratum. To compute the variance it is usual to imagine the implicit strata grouped in pairs to give 'pseudo-strata' and to regard each pair of adjacent selections as having been drawn independently from a single pseudo-stratum. This procedure, called the *collapsed strata technique*, may overestimate the variance of the systematic sample. In most practical situations, the overestimation is unimportant and is preferable to sacrificing the convenience (and probably somewhat lower variance) of systematic sampling.

[2] That is, selection from a list at a fixed interval, with a random start.

Often systematic sampling is done within explicit strata. In this situation, the pseudo-strata formed by collapsing implicit strata should be created in such a way as not to cut across the original explicit strata. Pairing of implicit strata will create a problem if there is an odd number to be dealt with; in this case one of the pseudo-strata can be made with three PSUs. Other variants of the above scheme are possible. In place of comparing L PSUs in L/2 pairs, the variance may be estimated by taking L-1 successive differences among the L units (ie taking the difference between PSUs 1 and 2, between 2 and 3, between 3 and 4, etc).

(3) Single Primary Selection per Stratum

To achieve an efficient sampling design, explicit stratification is sometimes carried to a point where only one PSU per stratum is selected into the sample. The Malaysia Family and Fertility Survey provides such an example among WFS surveys: here the rural sector was divided into 70 strata and one PSU per stratum was selected.

The situation in this case is similar to that in systematic sampling. An exact measure of variance must be abandoned in favour of an approximation based on the collapsed strata technique, in which the actual strata may be paired so that each of the resulting pseudo-strata is assumed to have a pair of independent selections. (Alternatively, one may use L-1 linked comparisons among L PSUs.) To reduce the overestimation of the variance involved, one should pair strata which are most alike. This pairing has to be done on the basis of likeness between *strata*; and not between the PSUs which happen to be selected: otherwise the variance may be seriously underestimated. This situation differs from pairing in a systematic sample in that no obvious criterion (such as position in an ordered list) may exist for determining the most appropriate pairing, thus requiring explicit examination of characteristics of individual strata. For this purpose information on these characteristics, perhaps on the basis of criteria used for stratification, must be compiled at the time of sampling and preserved. This is by no means an automatic process, as was demonstrated by the difficulties experienced in computing sampling errors for the above-mentioned survey in Malaysia.

(4) Grouping of PSUs

In samples involving large numbers of PSUs with small samples selected per PSU, one may, for convenience and economy, group PSUs appropriately to form pseudo-PSUs for the purpose of computing sampling errors. If the grouping of PSUs is done on a random basis (within each stratum separately), the overestimation of the variance involved in the above procedure is kept small.[3] Two examples of this technique applied to WFS samples are the following.

(a) In certain cases such as Nepal and Malaysia, although the *rural* sample employed a multi-stage design, in the *urban* sector households were selected directly using single-stage random or systematic sampling. To compute sampling errors for the total sample in a convenient manner, the urban sample households could be grouped *randomly* into pseudo-clusters similar in size to the actual clusters in the rural sector.

---

[3] This procedure is similar to the following. Suppose that the units in a simple random sample are grouped at random into 'pseudo-PSUs'. Variance for the actual random sample may be estimated reasonably well by treating the sample as if it were a clustered sample with the random groups of ultimate units as actual PSUs. This is because the 'clusters' formed by random grouping of elements are expected to have zero intracluster correlation (see p 29 below) so that the expected value of the computed variance is approximately the same as that for the simple random sample.

(b) The sample for the Sri Lanka Fertility Survey employed a large number (750) of clusters many of which were very small (a few actually contained no completed interviews). The clusters containing fewer than five interviews were grouped together to yield a total of 606 pseudo-clusters which were then used for variance computation.

(5) Effective Sampling Stages

It is useful to clarify certain ambiguities which may arise in the concept of 'sampling stages', and which concern the procedures for computing sampling errors. Multi-stage sampling or clustering is introduced to save costs of (a) travel and supervision, and (b) sample frame construction, listing and sample selection. The second consideration may be particularly important in situations where the available sampling frame does not provide area units of sufficiently small size for efficient (cost effective) sample design. For example, it has been a common practice in WFS surveys (Turkey, Indonesia, Senegal, Syria) to select area units in a number of 'steps' proceeding from larger to smaller categories of units, but not in such a way as to produce an additional clustering of the smaller units. The objective of introducing these steps was to limit the work necessary for constructing the sampling frame. Any clustering of the smaller units was avoided by selecting only one unit from each larger unit selected (Verma 1981a). For example, in one sampling domain of the Turkish Fertility Survey, a number of localities (towns) were selected with probability proportional to size (PPS); within each selected locality, ward maps and population data were compiled and updated where necessary, and *one* ward was selected with PPS; each ward was mapped in greater detail, divided into segments and *one* segment selected with PPS; finally within each selected ward households were listed and sampled with appropriate probabilities (Hacettepe Institute of Population Studies 1980). The procedure reduced enormously – in fact, made feasible – the work necessary to construct a frame of segments, but the expected sample outcome in no way differed from what it would have been had segments been selected directly from a frame of segments (actually non-existent). The sample described above is effectively a two-stage sample, at least as concerns the sampling errors: segments being the first or primary sampling units, and households the second or ultimate stage units. The earlier stages or steps not resulting in additional clustering of the sample are not relevant in the context of sampling error estimation.

(6) 'Self-Representing' Units

The term 'self-representing PSUs' is sometimes used to refer to area units which appear in the sample with certainty. This situation has arisen in several WFS surveys, particularly in the Latin American region. It usually happens when a certain type of unit referred to in the sample design is much larger in population size in one sampling domain (say the urban sector) compared to the same type of unit in another (say the rural domain). Consider, for example, a multi-stage design in which counties form the first-stage units to be selected with PPS, districts within counties form the second-stage units, and households within districts form the third or ultimate stage units. Suppose, however, that some of the counties are so large that they are taken into the sample with certainty and that within each a sample of districts is selected directly. The description of such counties as 'self-representing PSUs', though common, is misleading and should be avoided. It is more appropriate, and necessary for sampling error computation, to describe each such county as a *stratum*, from which (in the above example) a two-stage sample is selected, with districts as the PSUs, and households as the SSUs. The other, 'non-self-representing' counties belong to the sampling domain with a three-stage design: counties as PSUs, districts as SSUs and households as the ultimate stage units.

(7) Coding of Sample Structure

Appropriate coding of the sample structure, preferably on the micro-level data files resulting from the survey, is an essential requirement to ensure that sampling errors can be computed properly, taking into account the actual sample design. This is not always done, as for example Kish *et al* (1976) found in their attempts to compute sampling errors for archived survey data in the United States. Information on strata, PSUs and sample weights, etc should be defined and coded *in the form required for sampling error computation.* For the computational procedures described in this section, this would require the following as a minimum.

(a) Identification of PSUs as they are to be used for computation, taking into consideration the points made in (4) to (6) above.

(b) Definition of effective strata, ensuring that at least two PSUs are present in each stratum. In samples with systematic selection of PSUs, or when only one PSU is selected per stratum, appropriate 'collapsing' or pairing of strata would be required, as explained in (2)-(3) above. In such a case, the pseudo-strata so defined should be separately identified and coded.

(c) Information on sample weights, if applicable.

Information on various sampling stages (units selected, sampling fractions, etc) will be necessary if the total sampling variance is to be decomposed into its components according to sampling stage.

3.4 SAMPLING ERRORS FOR COMPLEX STATISTICS

The procedures outlined in section 3.2 fall in the class of methods called 'Taylor expansion method'. This is the approach followed in the WFS package program CLUSTERS for computing sampling errors for ratios and certain differences of ratios (Verma and Pearce 1978; see annex I).

Numerical procedures for computing variances of other more complex statistics using the Taylor expansion method have been developed (see for example Tepping 1968; Woodruff 1971; Woodruff and Causey 1976). Basically the method produces an estimate of the variance of a statistic based on variances of the linear (first order) terms of the Taylor series expansion of the statistic. Suppose we wish to estimate the variance of an estimator z which itself is a function of linear estimators (such as sample totals) z_k. Then it can be shown that, if the sample is sufficiently large for the Taylor approximation to be valid, the variance of z is approximated by the variance of a *linear* combination of the z_k's, ie

$$\mathrm{var}(z) = \mathrm{var}\left(\sum_k d_k \cdot z_k\right) \tag{3.17}$$

in which the d_k are the partial derivatives of z with respect to z_k, $d_k = \partial z/\partial z_k$, and are treated as *constants* in equation 3.17.

As an illustration, consider the ratio $z = z_1/z_2$ of two sample totals z_1 and z_2. We have

$$d_1 = \frac{\partial z}{\partial z_1} = \frac{1}{z_2}; \qquad d_2 = \frac{\partial z}{\partial z_2} = -\frac{z_1}{z_2^2} = -\frac{z}{z_2}$$

so that, from equation 3.17,

$$\text{var}(z) = \text{var}(d_1 z_1 + d_2 z_2) = d_1^2 \text{ var}(z_1) + d_2^2 \text{ var}(z_2) + 2d_1 d_2 \cdot \text{cov}(z_1, z_2)$$

$$= \frac{1}{z_2^2} [\text{var}(z_1) + z^2 \cdot \text{var}(z_2) - 2z \cdot \text{cov}(z_1, z_2)].$$

This is the same expression as equation 3.6 above, with $z_1 = y$ and $z_2 = x$.

The procedure for computing var(z) is considerably simplified by introducing the variable (Woodruff 1971):

$$z_{hij} = \sum_k d_k \cdot z_{k,hij} \tag{3.18}$$

where, as before, h, i, j stand for stratum, PSU and ultimate unit respectively. For example, for the ratio $z = z_1/z_2$ considered in the above illustration

$$z_{hij} = \frac{1}{z_2} (z_{1,hij} - z \cdot z_{2,hij})$$

which can be seen to be identical to equation 3.11, with $z_1 = y$ and $z_2 = x$.

Similarly one can easily derive equations 3.13–3.16. Equation 3.18 also provides the basis for estimating sampling errors of complex statistics in complex samples, such as coefficients in a regression equation.

Two other commonly used methods for estimating variances of complex statistics are the balanced repeated replications (BRR) and jackknife repeated replications (JRR) methods. These methods are based on the concept of replications described in section 3.1. Essentially, with the BRR method a replication consists of a random half of the total sample, and estimates the variance of the entire sample. With the JRR method, a replication is made up of a random half of one stratum plus the rest of the sample; and consequently, each replication measures the variance contributed by a single stratum. Empirical illustrations of the use of these methods for computing sampling errors for complex statistics from complex samples are given in Kish and Frankel (1974).

3.5 VARIABILITY OF VARIANCE ESTIMATES

It is important to realize that variance estimates from a sample are themselves subject to variability, particularly for samples based on relatively small number of PSUs. As noted by Kish *et al* (1976: 19), 'sampling theory, and experience with many and repeated computations, teach us not to rely on the precision of individual results, even when these are based on samples with a large number of elements'.

The precision of variance estimates is a complex subject. For reasonably large samples with good control to eliminate extreme variations in cluster size, a useful approximation to the coefficient of variation of a variance estimate is (Kish 1965: 289–91):

$$cv^2 = 2/df,$$

where df is the degrees of freedom, approximately equalling the number of PSUs, less the number of strata. For example, when two PSUs are selected from each of H strata (the common paired selection model),

$$df = (2H - H) = H,$$

so that

$$cv = \sqrt{2/H}.$$

Thus, for a sample with 100 PSUs from 50 strata,

$$cv = \sqrt{2/50} = 0.2,$$

while with only 25 PSUs (from 12 or 13 strata)

$$cv = 0.4.$$

3.6 CONFOUNDING OF SAMPLING AND RESPONSE VARIANCES IN COMPUTED SAMPLING ERRORS

In estimating sampling variance from equations such as 3.2–3.4, the element values (y_{hij}, x_{hij}) are meant to be free from non-sampling variability, ie they are the expected values for particular elements obtained by all possible measurements under the same essential conditions (see section 2.2). In practice, a particular survey yields only a single set of observed values, resulting in some degree of confounding of sampling and non-sampling variance in the usual estimation of the former.

In explaining this confounding, it is useful to distinguish between two components of response (or other measurement) variance: correlated response errors and uncorrelated response errors. Each interviewer, supervisor or coder, etc may have his own bias, which affects all the interviews which make up his workload. In so far as individual survey workers have different average effects on their respective workloads, they introduce errors which are correlated for all interviews within their individual workloads. In addition to these correlated errors, there may be chance factors which affect responses obtained from individual respondents independently of the particular survey worker involved. These are uncorrelated response errors. (For a fuller account see Hansen *et al* 1961.)

From a single survey, there is no way to separate out the confounding of uncorrelated (or simple) response errors from the usual estimation of sampling errors. This would require at least two independent measurements on each respondent (for a description of methodology and application to WFS data, see O'Muircheartaigh and Markwardt 1981, and O'Muircheartaigh 1982).

The relation of correlated response errors to the usual estimates of sampling errors is more complex. In a sense, survey workers impose their own 'clustering' on the sample of observations because of their individual biases. In so far as this clustering coincides with the geographic clustering of the sample itself and different fieldworkers are employed in different PSUs in each stratum (as for example will be the case if fixed enumerators are used, one for each sample cluster), then the usual estimate of sampling error fully includes the contribution of correlated response errors due to interviewer bias. The situation with most WFS surveys is rather different. Usually interviewers are organized in teams of four or five who share work in each PSU, and each team completes fieldwork in a number of PSUs, usually covering all PSUs in a stratum. With such an arrangement, the estimated sampling errors in the main do not include the contribution of correlated response variance due to the interviewer effect. Simple response variance is of course included as always.

4 Patterns of Variation and Portability

4.1 OBJECTIVES OF INVESTIGATION

For a number of reasons it is useful to investigate the patterns of variation of sampling error results across variables, across sample subclasses and across surveys, and to relate these patterns to the structure of the sample. These reasons are discussed below.

(1) Extrapolation of Computed Results

Generally the estimates of interest from a large-scale multi-purpose survey are too numerous for sampling errors to be computed for all of them. For example, the detailed cross-tabulations recommended for WFS surveys run into thousands of cells (WFS 1977). Ideally, the user of survey results needs to be able to obtain at least an approximate value of the standard error not only for the estimate in any cell of the detailed tabulations, but also for differences and distributions across cells. This can only be achieved by providing some means of extrapolation of errors from computations for selected variables and sample categories, to other variables and categories for which actual computation was not performed. This requires a study of the patterns of variation of sampling errors across variables and subclasses.

This may be particularly relevant when all survey estimates have to be reproduced for a number of reporting domains. Examples are WFS surveys in Fiji, where the entire set of tabulations is repeated for two ethnic groups and of course for the total sample; and Turkey and Indonesia where a substantial number of tables are repeated by region and type of place of residence. In view of the greatly increased number of survey estimates involved, extrapolation of sampling error results across major reporting domains becomes necessary.

A similar consideration is often involved in repetitive surveys with the same or similar design and content. Under such conditions, the standard error or some statistic derived from it may be relatively stable from one survey to the next, so that once the variance pattern is established in the beginning, it can be utilized to predict sampling errors for subsequent rounds.

(2) Summarization for Presentation

While it is desirable to provide the user of survey results with all the required information on sampling errors, it is necessary to do so in a way that is convenient for the user and that does not obscure the substantive results, which are after all of primary interest.

This presents a similar problem to that discussed in (1). Even if the sampling errors for all the published estimates were computed, it would not be feasible to publish them: they would double the volume of the tables in the report, even without considering the sampling errors of *differences*. It is essential to provide the user with some simple way of computing approximate sampling errors for any estimates and differences in which he may be interested. Once again, this implies some means of extrapolation and this, in turn, requires a knowledge of the patterns of variation of sampling errors.

(3) Smoothing of Computed Results

As noted earlier, sampling errors computed from sample data are themselves subject to considerable variability, particularly for samples based on relatively small numbers

of PSUs. In fact it may often be *preferable* to use results appropriately averaged over a number of computations, rather than to rely on the precision of individual computations.

(4) Sample Design and Evaluation

Apart from indicating the precision of existing survey estimates, the objective of sampling error computation is to evaluate how a particular design has fared and to provide data for designing future samples. For this, it is necessary to explore patterns of variation of sampling errors as related to important features of the sample structure, such as clustering, stratification and weighting. In fact, as will be seen later, the relationship between sampling errors and sample structure is useful also in the extrapolation, summarization and smoothing of computed results within a given survey.

For the various reasons stated above, it is necessary to combine somehow the results from computations for different variables and subclasses on the basis of which patterns of variation can be established more clearly. Pooling of results for different variables is more problematic (Kish *et al* 1976), but perhaps also less critical since the number of variables involved is usually substantially smaller than the number of subclasses of interest, since the sample can be divided into subgroups in numerous ways. In any case, it is important to recognize that, while smoothing, pooling and extrapolation of computed sampling errors is often necessary and desirable, there are risks involved in doing this. Excessive or careless application of these procedures can hide actual variation, distort the result and mislead the user. The only guarantee against this is to base extrapolation and smoothing on *an extensive and wide variety of actual computations*, covering many variables and subclasses of different types, and to check how well the smoothed results fit the actual computations.

4.2 PORTABILITY

The Concept

To meet the requirements for extrapolation, summarization and smoothing of computed results, it is necessary to search for portable measures of sampling variability. The term 'portability', introduced by Kish, refers to the possibility of carrying over from one subclass to another, from one variable to another or from one survey to another, the conclusions drawn regarding the sampling error. To illustrate the concept, suppose that a number of self-weighting simple random samples (SRS) of different sizes are drawn from a population to measure the same set of variables. The variance of an estimated mean $\bar{y}$ for a sample of size n is $\frac{1-f}{n}\sigma^2$, which is estimated by

$$sr^2(\bar{y}) = \frac{1-f}{n}\left[\sum_j \frac{(y_j-\bar{y})^2}{n-1}\right] = \frac{1-f}{n}\cdot s^2, \text{ say} \qquad (4.1)$$

While the standard error varies inversely as $n^{1/2}$, the standard deviation, σ, is the same for different samples. It measures the root mean square deviation of individual values y_j from the mean and is portable across samples.

Different measures are portable to different degrees. The standard error is specific to the estimate for which it is computed, and its magnitude depends upon a number of factors such as:

1 The nature of the estimate;

2 its units of measurement (scale) and magnitude;

3 its variability in the population;

4 the sample size;

5 the sample design (clustering, stratification, weighting, cluster size, etc);

6 the nature and size of the sampling units;

7 for sample subclasses, their nature and distribution across sample clusters.

Standard errors (se) computed for one statistic can, at best, be imputed directly only to essentially similar statistics, based on samples of similar size and design. Various derived measures of se are introduced to control, ie to reduce, the effect of some of the above factors, and hence enhance the portability of the measure across subclasses, variables and sample designs. In the example given above, the effect of sample size (n) was controlled by introducing σ in place of the actual se.

Zarcovich (1979) illustrates in detail with practical examples how the estimated coefficient of variation (standard error divided by the mean, $se(\bar{y})/\bar{y}$) can be stable across a number of repetitive surveys with similar design, size and content, thus eliminating the need for fresh calculations in each survey round. This measure controls for factor 2 above, namely for units of measurement and magnitude of the estimate.

Two more useful and widely used measures of portability are the design factor (deft) and rate of homogeneity (roh) described below.

Design Factor (deft)

An extremely useful measure in this connection is the design factor (deft),[4] defined as the ratio of the estimated standard error for the actual design (se) to the estimated standard error for a simple random sample (SRS) of the same size (sr):

$$\text{deft} = \text{se}/\text{sr} \tag{4.2}$$

This measure is more portable than se, since it does not depend upon factors which affect both se and sr in the same way, factors such as units of measurement, magnitude of the estimate, its variability in the population, and above all, sample size. Deft depends upon other factors such as the nature of the estimator, sample design, and type and size of sampling units. Deft is a summary measure of the effects of departure of the actual sample design from SRS. It is a comprehensive factor which attempts to summarize the effect of various complexities in the design, especially those of clustering and stratification. It may include even the effect of ratio or regression estimation, of double sampling and of varied sampling fractions. For these reasons many samplers include the ratio se/sr as a routine item in the output of variance computations.

To estimate deft from a sample it is necessary to estimate both se and sr. As described in section 3, se for multi-stage complex sample designs can, in many practical situations, be estimated simply from quantities aggregated at the level of PSUs, without explicit

[4] We use the term 'design factor' to mean the ratio of actual standard error to SRS standard error; the term 'design effect' (deff) is normally used for the ratio of the variances.

$$\text{deff} = \text{deft}^2 = \text{se}^2/\text{sr}^2$$

reference to subsampling procedures within PSUs. An equally convenient result of sampling theory is that, in many practical situations, the sampling error corresponding to an SRS can be estimated from a complex sample simply by ignoring the complexity of the actual design. For example, for the ratio r defined in equation 3.5 the SRS variance is approximately

$$sr^2(r) = \frac{1}{n-1} \cdot \Sigma(w_{hij} \cdot z_{hij}^2)/\Sigma w_{hij}, \text{ with } z_{hij} = y_{hij} - r \cdot x_{hij} \qquad (4.3)^5$$

even though the actual observations (y_{hij}, x_{hij}) are from a complex sample rather than from an SRS.

Rate of Homogeneity (roh)

For a given variable and a given number and type of clusters and subsampling procedure used, the value of deft tends to increase with increasing cluster size. To control this effect, Kish (1965) introduced a synthetic measure roh (rate of homogeneity) defined as

$$deft^2 = 1 + (\bar{b} - 1)\, roh \qquad (4.4)$$

where $\bar{b}$ is the average cluster size. The model is based on the concept of intraclass correlation which measures the degree of correlation between members of a cluster. Equation 4.4 has been developed for self-weighting samples in the absence of extreme variation in cluster sizes. Roh is a synthetic measure introduced with the aim of measuring the average degree to which values of a particular variable are homogeneous within PSUs, relative to that variable's overall variability.

The following illustration may clarify the relationship of deft and roh to cluster size $\bar{b}$. Suppose that a two-stage sample of size n = 2500 is drawn by selecting 49 clusters, and by selecting at random an average of $\bar{b} = 51$ ultimate units per cluster. Assume that for a particular variable $deft^2 = 2$ for this sample; in other words, the variance of the clustered sample is twice as large as that of an SRS of the same size. The implication is that an SRS of size $n' = n/deft^2 = 1250$ would have given the same sampling precision. (It is important to realize that the above statement is true only for estimates based on the total sample; as discussed later, defts can be much smaller for subclasses.) The implied value of roh for the variable is

$$roh = \frac{deft^2 - 1}{\bar{b} - 1} = \frac{1}{50} = 0.02$$

Now suppose that, retaining the same number of clusters, the average sample per cluster is reduced to $\bar{b} = 26$. One expects roh to be unchanged (ie it is portable between the two samples) because the nature of the sampling units and the sampling procedure have not changed. The design effect becomes

$$deft^2 = 1 + 0.02\,(26 - 1) = 1.5$$

In the above sense, roh removes the effect of $\bar{b}$ in deft, and is a more portable measure. However, it should be emphasized that roh is specific to a particular variable, sample design and type of sampling unit. Note that for the simple random sample, halving the sample size doubles the sampling variance. For the clustered sample in this example,

---

[5] The relative bias involved in this procedure is approximately $(deft^2 - 1)/n$, which is negligible in most situations with reasonably large sample size. In fact this procedure can be used to estimate the effect on sampling error of particular features such as stratification or additional area stages, by repeating the calculation ignoring that particular feature of the design. Illustrations are provided in Verma *et al* (1980).

halving the sample size (but retaining the same number of clusters) increases the variance by a factor of only 1.5.

4.3 MODELLING OF SAMPLING ERRORS FOR SUBCLASS MEANS

The measures deft and roh provide empirically useful means of modelling sampling errors across diverse subclasses of a sample. In this section we consider how, for a given substantive variable or group of similar variables, standard errors of ratios, means and proportions for *subclasses* (se_s) may be related to those for the *total* sample (se_t). When considering subclasses instead of the entire sample there are three important points to bear in mind:

1 In so far as a subclass is spread evenly over all sample clusters, the effective cluster size (ie sample size per PSU) is reduced compared with that for the total sample, the reduced figure being roughly proportional to the subclass size.

2 However, this is not the case for subclasses which are confined to a subset of sample clusters. Furthermore, the estimates of variance tend to be less stable in this case since they are based on only a subset of the PSUs.

3 In any case, subclasses are rarely uniformly distributed, so that the coefficient of variation of cluster size tends to be higher for a subclass than for the total sample. This would tend to increase not only the error variance but also the bias in ratio estimation.

The variation of sampling error with subclass size is therefore related to how evenly the subclass is distributed over sample clusters. In this respect it is useful to distinguish three types of subclasses (Kish *et al* 1976). First, certain classes such as groups defined in terms of demographic characteristics (age, sex, marriage duration, etc) tend to be more or less uniformly distributed geographically across the whole population, and hence across the sample clusters. These may be called *cross-classes.* At the opposite end we have *geographic classes* which are completely segregated into separate clusters, ie a whole cluster either belongs or does not belong to the subclass. Examples are regions, or urban-rural domains, of a country. Other classes, such as particular ethnic, occupational or other socio-economic groupings, while less well distributed than cross-classes, are not as completely segregated as geographic classes. For example, higher educational groups, and even more so, non-farming occupations tend to be clustered in, though not confined to, urban areas. These are termed *mixed classes.*

Sampling Errors for Cross-Classes

For a given variable we may expect the subclass design factor, $deft_s$, to be smaller than total sample $deft_t$, since the effective cluster size decreases proportionately with decreasing cross-class size. For small cross-classes in a self-weighting sample, the effective sample would tend towards SRS, ie $deft_s$ would tend towards unity. Taking account of the above effect, one might attempt to give a more precise expression to the relationship between sampling errors for subclasses and the total sample by adopting a model based on equation 4.4. Thus

$$\frac{(deft_s^2 - 1)}{(deft_t^2 - 1)} = \frac{roh_s \cdot (\bar{b}_s - 1)}{roh_t \cdot (\bar{b}_t - 1)} \qquad (4.5)$$

For a cross-class we assume $roh_s = roh_t$ so that the right-hand side becomes

$$\frac{\bar{b}_s - 1}{\bar{b}_t - 1}$$

which may be approximated by M_s, the size of the subclass in proportion to the whole sample, provided that $\bar{b}_s$ is substantially greater than 1. Thus

$$deft_s^2 = 1 + M_s (deft_t^2 - 1) \tag{4.6}$$

According to the model the departure of $deft_s^2$ from unity is proportional to the size of the subclass relative to total sample size. For very small subclasses $deft_s$ tends to 1.0, ie the effective sample tends to SRS.

The model needs to be modified for samples which are not self-weighting. While the effects of clustering and stratification tend to disappear for very small cross-classes, the effect of sample weighting tends to persist. When population variances and sample weights are uncorrelated, deft will be found to be greater than 1.0 even for very small subclasses; the effect of unequal weights (uncorrelated with the population variance) is to multiply the variance of all estimates by the factor (Kish 1965)

$$L = \frac{\Sigma n_h w_h^2 \cdot \Sigma n_h}{(\Sigma n_h \cdot w_h)^2}, \tag{4.7}$$

where n_h is the number of units with weight w_h. It is found in practice that deft for very small cross-classes (and also for differences between such subclasses, see section 4.4) tends to the value $L^{1/2}$ in accordance with equation 4.7. Table 1, from Verma *et al* (1980), demonstrates this on the basis of a very large number of computations. The design factors shown are averaged values over groups of similar variables. These groups cover most of the estimates of substantive interest from the WFS individual questionnaire. The groups are: (a) seven variables concerning *nuptiality*, such as age at marriage, marital and exposure status, marriage dissolution and re-marriage; (b) eleven *fertility* and related variables, such as number of children ever born, the number currently living, measures of fertility in specified periods, birth intervals, duration of breast-feeding; (c) six variables concerning fertility preferences, such as son preference, the desire to stop childbearing, the additional number of children wanted and the total desired family size; (d) four variables concerning the *knowledge* of various methods of contraception; and (e) ten variables concerning contraceptive *use*, by specified method and timing of use. The row 'effect of weighting' is $L^{1/2}$ computed from equation 4.7 (equation 6.1 in the source being quoted), and the agreement between this and the computed deft is very close indeed. Thus it is convenient to define an adjusted design factor deft' *excluding* the effect of weighting equation 4.7, ie as

$$deft = se/sr = L^{1/2} \cdot deft' \tag{4.8}$$

This allows an expression of the form of equation 4.6 to relate $deft_s$ to $deft_t$ for non-self-weighting samples also. Hence we can write standard errors for the total sample (se_t) and for a subclass (se_s) as:

$$se_t^2 = \frac{s_t^2}{n} \cdot L_t \cdot deft_t'^2 \quad \text{for the total sample,}$$

and

$$se_s^2 = \frac{s_s^2}{n_s} \cdot L_s \cdot deft_s'^2 \quad \text{for the subclass.} \tag{4.9}$$

To relate se_s to se_t, we relate the subclass value to the total sample value for each of the three quantities on the right-hand side of equation 4.9.

Table 1 Deft values for small selected 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[a]			Rural	Total
Effects of weighting[b]	1.06			1.12			1.18			1.19			1.09			1.11			1.00	1.06
Subclass/difference[c]	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(1)
Subclass $\bar{b}$	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[d]																				
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	1.05	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

[a] Includes the small 'estate' domain.
[b] Increase in standard error due to departure from self-weighting within country or domain estimated from equation 4.7 (equation 6.1 in Verma *et al* 1980).
[c] (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'.
[d] For fuller details, see text p 31.
NOTE: For differences, b is defined as half the harmonic mean of $\bar{b}_s$ for the two subclasses.
Source: Verma *et al* (1980)

For a cross-class, well distributed over the population and hence over sample clusters, we have the following relationships:

1 The deft values are related by equation 4.6, namely

$$\text{deft}_s'^2 = 1 + (\text{deft}_t'^2 - 1) \cdot M_s \tag{4.10}$$

where deft′ is defined by equation 4.8.

2 It is reasonable to assume

$$L_s = L_t \tag{4.11}$$

since the relative allocation among domains (n_h in equation 4.7) for the subclass will in general be similar to that for the total sample. Table 1 demonstrates this.

3 However the values of standard deviations, s_s and s_t, may differ considerably for the following reasons. Subclasses of interest usually correspond to cross-classifications introduced to control for factors correlated with the substantive variable being estimated. For example, to compare mean fertility among different educational groups, data may be classified by age, so as to control for age differences in the educational groups being compared. Due to the strong relationship between age and fertility, subclasses defined in terms of age may divide the sample into somewhat more homogeneous groups of fertility than the total sample. More importantly, the mean value of the fertility measure could differ appreciably from one age group to another.

Generally, s_s and s_t may be related on the basis of their relationship to their respective means. For a dichotomous response leading to cross-classification of proportions (p), we have the well-known binomial expression, namely

$$s^2 \propto p(1 - p) \tag{4.12}$$

For means (m), it may be reasonable to assume the Poisson distribution

$$s^2 \propto m \tag{4.13}$$

Little (1978) notes that the above distribution is particularly appropriate for responses which are accumulated counts (such as cumulative fertility, ie children ever born). He also suggests the following more general relationship.

$$s^2 \propto m^{\alpha} \tag{4.14}$$

where the exponent α may be estimated empirically by fitting the model to actual computations. With $\alpha > 1$, the variance would increase more rapidly with m than the Poisson distribution (equation 4.13); and conversely for $\alpha < 1$. $\alpha = 0$ corresponds to the assumption of homoscedasticity, ie s^2 = constant and hence $s_s^2 = s_t^2$ in equation 4.9.

To provide an illustration of the ideas developed so far, table 2 shows standard errors for a number of substantive variables for the total sample of women and for a number of subclasses defined as age groups from the Turkish Fertility Survey (Hacettepe Institute of Population Studies 1980). A variety of subclass sizes are covered, and some of the subclasses are partly overlapping. For each variable, two rows are shown: (a) actually computed subclass variances, se_s^2; and (b) se_s^2 predicted from total sample se_t^2 using equation 4.9 on the assumptions that

- deft_s' and deft_t' are related by equation 4.10, since the age classes are more or less true cross-classes;

Table 2 Comparison of (a) computed and (b) predicted[a] subclass standard errors for selected variables from the Turkish Fertility Survey

Subclass (age group)	15–19	15–24	20–24	15–34	25–29	25–34	30–34	35–39	35–44	35–49	40–44	45–49	Total
Sample size of subclass[b]	345	1156	811	2678	840	1522	682	644	1255	1753	611	498	4431
Means													
Age at first marriage													
(a)	[c]	–	–	0.095	0.106	0.095	0.144	0.132	0.099	0.088	0.122	0.141	0.072
(b)			–	0.091	0.113	0.091	0.124	0.126	0.097	0.086	0.129	0.142	–
Children ever born													
(a)	0.044	0.039	0.055	0.053	0.071	0.072	0.109	0.134	0.117	0.111	0.154	0.181	0.060
(b)	0.164	0.096	0.111	0.070	0.110	0.086	0.120	0.123	0.093	0.081	0.126	0.138	–
Total number of children desired													
(a)	0.093	0.050	0.055	0.043	0.049	0.052	0.079	0.111	0.079	0.067	0.076	0.087	0.045
(b)	0.096	0.061	0.067	0.049	0.068	0.056	0.074	0.077	0.061	0.056	0.079	0.089	–
Proportions													
Ever heard of pill													
(a)	0.025	0.015	0.016	0.009	0.013	0.010	0.015	0.018	0.016	0.015	0.020	0.023	0.010
(b)	0.023	0.014	0.016	0.011	0.016	0.013	0.017	0.018	0.014	0.013	0.018	0.020	–
Ever used any method of contraception													
(a)	0.024	0.017	0.020	0.015	0.020	0.018	0.024	0.022	0.019	0.016	0.025	0.023	0.013
(b)	0.029	0.018	0.021	0.014	0.020	0.017	0.022	0.022	0.018	0.016	0.023	0.025	–
Currently using a modern method of contraception													
(a)	0.019	0.012	0.015	0.010	0.016	0.013	0.022	0.016	0.012	0.012	0.019	0.026	0.008
(b)	0.024	0.013	0.016	0.009	0.015	0.011	0.016	0.016	0.013	0.012	0.019	0.026	–

[a] For basis of prediction, see p 33.
[b] The actual sample base is smaller for a variable which does not apply to all respondents (eg the variable on current use of contraception).
[c] Not defined.
Source: Hacettepe Institute of Population Studies (1980)

Table 3 Comparison between computed and predicted subclass standard errors for the estimated mean number of children ever born, by age group of women

Age group	15–19	15–24	20–24	15–34	25–29	25–34	30–34	35–39	35–44	35–49	40–44	45–49
Mean children ever born (m)	0.670	1.469	1.809	2.663	2.991	3.570	4.283	5.483	5.713	5.881	5.956	6.303
Predicted se assuming $s_s = s_t$	0.164	0.096	0.111	0.070	0.110	0.086	0.120	0.123	0.093	0.081	0.126	0.138
Predicted se assuming $s \propto m$	0.068	0.059	0.076	0.058	0.096	0.081	0.125	0.146	0.112	0.100	0.155	0.180
Computed se	0.044	0.039	0.055	0.053	0.071	0.072	0.109	0.134	0.117	0.111	0.154	0.181

Source: Hacettepe Institute of Population Studies (1980)

- $L_s^2 = L_t$ (which is defined to be 1 since the sample is self-weighting);
- $s_s^2 = s_t^2$, ie standard deviations are *assumed* equal.

The agreement between computed and predicted se_s^2 is generally good, except in the following cases:

- for the variable 'children ever born', the predicted values are substantially higher than the computed values for younger age groups (and therefore for women with lower mean number of children ever born), the opposite is the case for the older age-groups;
- for the three proportions shown in the table, the discrepancy is generally small except for the age group 15–19, that is, for the youngest women in the sample.

These discrepancies are related to the assumption $s_s^2 = s_t^2$ made in table 2. There is a notable variation in the mean number of children ever born among the age groups, the mean increasing from 0.67 for the youngest group (aged 15–19) to 6.30 for the oldest (45–49). Assuming $s^2 \propto m$ (equation 4.13), the agreement between computed and predicted values is greatly improved, as shown in table 3. The agreement can be further improved by fitting equation 4.14 or a relation of the form

$$s^2 = \alpha + \beta . m \tag{4.15}$$

There is little variation by age in the mean of the other two variables in equation 4.2, so that introducing equations 4.13–4.15 does not make much difference.

Regarding the proportions in table 2, the 'correction factor' $p(1 - p)$, from equation 4.12, is rather insensitive to the value of p in a broad range around $p = 0.5$. The proportions differ significantly from the overall values only for the youngest age group; using equation 4.12, the agreement is considerably improved, as shown in table 4.

Mixed Classes

In so far as subclasses are unevenly distributed across sample clusters, the coefficient of variation of effective cluster size, and hence the variance (and bias) of ratio estimators, will tend to increase. Consequently the design factor for mixed classes is expected to decline less rapidly with decreasing subclass size than it does for the well-distributed cross-classes. In a study of the pattern of variation of design factor with subclass size, Verma *et al* (1980) propose the following model, generalized from equation 4.10 to fit mixed classes:

$$deft_s^2 = 1 + (deft_t^2 - 1) . M^{\alpha} \tag{4.16}$$

where α is an empirically determined parameter expected to be in the range 0 to 1, with values at the upper end corresponding to cross-classes, and at the lower end to segregated

Table 4 Standard errors for subclass aged 15–19: comparison between computed and predicted values for proportions

	Predicted se_s		Computed
	Assuming $s_s = s_t$	Assuming $s \propto p(1 - p)$	
Proportion ever used contraception	0.029	0.024	0.024
Currently using modern method	0.024	0.018	0.019

or geographic classes. The model was tested on the basis of sampling errors computed for different types of variables over many subclasses from WFS surveys in a number of countries. This is a useful model for the survey statistician confronted with the task of summarizing and extrapolating sampling errors for subclasses, and it may be helpful here to sketch the procedure used for estimating α and to reproduce some of the results from the above-mentioned study:

The simple model proposed above assumes that, for a given degree of cross-classedness, the substantive nature of the variable as well as that of the characteristic defining the subclass need not be considered. This, however, is most unlikely to be the case, and thus we estimated α separately for similar groups of variables within subclass groups of similar type and cross-classedness within each country. The estimation procedure is to minimize the sum of squared deviations of the fitted from the observed value of deft_s.

Empirically we found that the goodness of fit improved considerably when the individual points were weighted by their relative sample sizes, namely M_s; since the error in the estimation of deft_s is inversely related to the sample size on which it is based, this transformation has the additional advantage of producing a more nearly homoscedastic distribution for the disturbance (or error) terms.

Thus the linear form of equation 4.16 from which α is estimated is

$$y_s = (1 + \alpha)x_s \quad \text{with} \quad y_s = -\ln(M_s \cdot d_s);\ x_s = -\ln(M_s), \tag{4.17}$$

where

$$d_s = (\text{deft}_s'^2 - 1)/(\text{deft}_t'^2 - 1).$$

The straight line, equation 4.17, is forced through the origin which corresponds to the total sample ($M_s = 1$), giving

$$1 + \alpha = \Sigma y_s \cdot x_s / \Sigma x_s^2$$

and a measure of goodness of fit

$$R^2 = 1 - \Sigma[(1 + d_s)x_s - y_s]^2 / \Sigma(y_s - \bar{y})^2.$$

Table 5, columns (5)–(6) show the measure R^2 and parameter α. Column (4) shows the number of cases (a 'case' = a variable estimated over a particular subclass) on which the estimate is based; as can be seen from this column, the estimation of α is based on a very large set of sampling error computations. The fit is reasonably good: in over 80 per cent of the sets, R^2 exceeds 0.5, and for over 50 per cent, $R^2 > 0.65$. The groups of variables in table 5 are the same as described earlier for table 1. Two groups of subclasses are considered: demographic subclasses (age, marriage duration, etc) which are more nearly cross-classes; and socio-economic subclasses (groups by level of education, occupation, etc) which are mixed classes. As expected, α values are larger for the more well-distributed demographic subclasses.

To relate the standard errors se_s and se_t, we refer back to equation 4.9. The standard deviations s_s and s_t are related as already described by equations 4.12–4.15; and often it is reasonable to assume $L_s = L_t$ even for ill-distributed mixed classes, since the effect of weighting equation 4.7 is not sensitive to moderate changes in the n_h values. However, for classes tending to be rather segregated (eg higher educational groups, usually concentrated in urban areas), it may be necessary to determine L_s by using more appropriate n_h values in equation 4.7.

Geographic Classes

Subclasses or domains completely segregated into separate clusters and strata present no special problems, since one may compute sampling errors for each domain separately

Table 5 Pattern of results for subclasses and subclass differences, by country, variable group and subclass group, averaged over selected variables and subclasses

			Subclass results						Subclass differences		
Country	Variable group[a]	$deft_t$ (1)	Subclass type[b]	$deft_s$ (2)	$\frac{deft_s^2 - 1}{deft_t^2 - 1}$ (3)	n_α (4)	R^2 (5)	α (6)	$deft_d$ (7)	$\frac{deft_d^2 - 1}{deft_s^2 - 1}$ (8)	β (9)
Mexico	Nuptiality	1.39	Demo.	1.15	0.35	52	0.59	0.76	1.05	0.32	0.91
			Socio.	1.15	0.35	72	0.49	0.47	1.13	0.86	0.96
	Fertility	1.58	Demo.	1.25	0.38	85	0.55	0.74	1.06	0.22	0.85
			Socio.	1.24	0.36	125	0.37	0.75	1.18	0.73	0.94
	Preferences	1.52	Demo.	1.20	0.34	33	0.66	0.67	1.08	0.38	0.88
			Socio.	1.22	0.37	45	0.39	0.73	1.18	0.80	0.93
	Knowledge	2.81	Demo.	1.79	0.32	31	0.85	0.87	1.08	0.08	0.60
			Socio.	1.73	0.29	36	0.77	0.85	1.46	0.57	0.82
	Use	1.92	Demo.	1.38	0.34	65	0.71	0.94	1.10	0.23	0.79
			Socio.	1.37	0.33	91	0.77	0.93	1.23	0.58	0.92
	All variables	1.70	Demo.	1.29	0.35	–	–	0.80	1.07	0.22	0.83
			Socio.	1.28	0.34	–	–	0.75	1.20	0.69	0.93
Thailand	Nuptiality	1.28	Demo.	1.21	0.73	28	0.71	0.45	1.10	0.45	0.98
			Socio.	1.20	0.69	37	0.50	0.21	1.09	0.43	0.98
	Fertility	1.38	Demo.	1.22	0.54	78	0.55	0.70	1.08	0.34	0.90
			Socio.	1.25	0.62	106	0.57	0.32	1.19	0.74	0.96
	Preferences	1.37	Demo.	1.25	0.64	39	0.32	0.16	1.07	0.26	0.88
			Socio.	1.22	0.56	57	0.53	0.12	1.20	0.90	0.95
	Knowledge	2.48	Demo.	1.60	0.30	30	0.68	1.14	1.02	0.03	0.65
			Socio.	1.61	0.31	37	0.77	0.77	1.69	1.17	0.95
	Use	2.15	Demo.	1.46	0.31	68	0.87	0.94	1.09	0.17	0.75
			Socio.	1.45	0.30	80	0.86	0.81	1.20	0.40	0.90
	All variables	1.65	Demo.	1.33	0.45	–	–	0.71	1.08	0.22	0.85
			Socio.	1.33	0.45	–	–	0.45	1.22	0.64	0.95
Bangladesh	Nuptiality	1.22	Demo.	1.12	0.52	24	0.75	0.99	1.06	0.49	0.95
			Socio.	1.18	0.80	33	0.76	0.46	1.17	0.94	0.99
	Fertility	1.12	Demo.	1.08	0.65	23	0.44	0.42	1.06	0.74	0.98
			Socio.	1.12	1.00	42	0.37	0.04	1.13	1.09	0.99
	Preferences	1.21	Demo.	1.11	0.50	29	0.49	0.49	1.04	0.35	0.94
			Socio.	1.19	0.90	53	0.57	0.16	1.19	1.00	1.00
	Knowledge	1.66	Demo.	1.28	0.36	35	0.81	0.93	1.07	0.23	0.84
			Socio.	1.45	0.63	49	0.63	0.47	1.32	0.67	0.91
	Use	1.31	Demo.	1.10	0.29	43	0.77	0.75	1.03	0.29	0.92
			Socio.	1.20	0.61	79	0.45	0.49	1.15	0.73	0.97
	All variables	1.26	Demo.	1.12	0.43	–	–	0.72	1.05	0.40	0.94
			Socio.	1.20	0.75	–	–	0.32	1.17	0.84	0.98
Indonesia	Nuptiality	1.45	Demo.	1.29	0.60	44	0.53	0.66	1.22	0.74	0.95
			Socio.	1.29	0.60	68	0.62	0.59	1.25	0.84	0.97
	Fertility	1.41	Demo.	1.29	0.67	69	0.35	0.46	1.21	0.70	0.94
			Socio.	1.29	0.67	112	0.62	0.44	1.25	0.84	0.98
	Preferences	1.55	Demo.	1.36	0.61	41	0.59	0.65	1.23	0.60	0.91
			Socio.	1.34	0.57	64	0.67	0.59	1.26	0.74	0.95
	Knowledge	2.44	Demo.	1.69	0.37	36	0.91	1.07	1.24	0.29	0.74
			Socio.	1.79	0.44	57	0.94	0.75	1.51	0.58	0.86
	Use	1.70	Demo.	1.37	0.46	74	0.61	0.85	1.20	0.50	0.87
			Socio.	1.45	0.58	115	0.79	0.64	1.36	0.77	0.95
	All variables	1.62	Demo.	1.37	0.54			0.74	1.21	0.53	0.90
			Socio.	1.39	0.57	–	–	0.60	1.31	0.77	0.95

[a] Same variables as table 1. For fuller details, see text p 31.
[b] See text, p 37. For specification of the subclasses used, see Verma *et al* (1980), appendix.
NOTE: $deft_s$ for subclasses, and $deft_t$ for total sample in this table are defined according to equation 4.2 and not equation 4.8, ie the effect of weighting has not been removed from the 'design factor' as defined here.
Source: Verma, Scott and O'Muircheartaigh (1980): 452

using the same method as for the total sample. However, it is convenient to relate results for domains to those for the total sample, not only to economize on computation and presentation but also because results for individual geographic domains tend to be less stable, since they are based on a smaller number of PSUs than the total sample.

In relation to the three components of equation 4.9 relating se_s and se_t:

1 Subclass and total sample standard deviations may be related as before, in accordance with equations 4.12–4.15.

2 The loss-factor due to weighting, L_s, for the domain may substantially differ from L_t for the total sample; often $L_s < L_t$ since weights are frequently introduced *between* domains with self-weighting samples within domains (a common example is the over-sampling of urban areas, with self-weighting sample ($L_s = 1$) within urban and rural domains separately).

3 In so far as the sample design and cluster sizes are similar between different geographic domains, we expect the same deft values. Generally, however, sample design may differ from domain to domain, and a simple model relating domain $deft_s'$ to total sample $deft_t'$ is

$$deft_s'^2 = 1 + c_s (deft_t'^2 - 1) \tag{4.18}$$

where c_s is a constant for the domain to be determined empirically by fitting the above relation to the computed $deft_s'$ over a group of variables.

Table 6 shows the results of fitting equation 4.18 to each of the 8 'type of place' domains from the Turkish Fertility Survey. The self-weighting sample consists of 215 PSUs, so that individual domain results are based on a small number (average 27, range 14 to 34) of PSUs, and hence have considerable variability. Each fit is based on sampling errors computed for 27 substantive variables covering a wide range of variable types. The goodness of fit R^2 varies from 0.1 to 0.6 with an average of around 0.3. Table 7 compares computed standard errors for selected individual variables, with predicted values on the basis of the least squares fit to equation 4.18 and the relationships of equations 4.12–4.13 applied to the pooled results for all variables.

Application to a Subclass Defined in Terms of Several Characteristics

In a multi-way cross-tabulation, a cell corresponds to a subclass defined in terms of a number of characteristics. Consider, for example, the mean number of children ever born, classified by women's age group, level of education and type of place of residence.

Table 6 Fitting of relation equation 4.18 to each of the eight domains by type of place of residence in the Turkish Fertility Survey

Domain, s	Metro-politan	Large cities	Medium cities	Small cities	Towns	Large villages	Medium villages	Small villages	Total
Average $deft_s'$ [a]	1.21	1.32	1.57	1.33	1.50	1.65	1.47	1.66	1.48
Estimated parameter c_s	0.39	0.64	1.25	0.64	1.05	1.47	0.98	1.48	–
Goodness of fit, R_s^2	0.11	0.26	0.37	0.28	0.19	0.56	0.22	0.29	–

[a] Averaged over 27 variables (covering all the 5 groups defined in table 1), computed results for which are used to fit equation 4.18 and thence to estimate c_s for each domain.
Source: Hacetepe Institute of Population Studies (1980)

Table 7 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_s)	648	697	350	318	628	497	734	559	4431
Cluster size, $\bar{b}_s$	19.1	19.4	21.9	22.7	20.3	19.1	21.6	24.3	20.7
No. of clusters a_s	34	36	16	14	31	26	34	23	214
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: Hacettepe Institute of Population Studies (1980)

To estimate the sampling error for, say, urban women, educated to the primary level and aged 25–29, one may proceed as follows:

1 Using the empirically fitted relation, equation 4.18, for geographic classes, determine the design factor for the variable 'children ever born' for the urban domain from the design factor for the total sample.

2 Apply equation 4.16, for mixed classes, *within* the urban domain to estimate deft for the class 'urban women, educated to the primary level'.

3 Within the above-mentioned class, apply equation 4.10 or 4.16 for cross-classes to obtain deft for the subclass of ultimate interest. Finally, se_s may be estimated using equation 4.9, with L_s corresponding to the urban domain and s_s^2 adjusted by equations 4.12–4.15 as appropriate.

As in the above example, it appears intuitively appropriate to proceed step by step from characteristics defining geographic classes to those defining mixed classes and finally to cross-classes.

The following is a simple illustration of the extrapolation procedure described in this section. The example is adapted from an actual computation.

Suppose that for a national sample of $n = 5{,}000$ women, the mean number of children ever born is $m = 4.02$, the computed design factor deft = 1.69 and the standard error se = 0.073. The sample is non-self-weighting with loss factor $L^{1/2} = 1.21$ (equation 4.7). Suppose that the objective is to estimate the standard error (se) for the subclass 'urban women who are literate and aged 20–29'. The sample weights are given to be less variable *within* the urban domain, with $L^{1/2} = 1.09$.

The observed means along with sample sizes for the relevant subclass are given in the first two columns of table 8. Figures *given* are in **bold** type. The remaining figures are computed from other data in the table. The values of c_s and α are assumed to have been estimated by fitting the models described to a set of actual computations of design factors.

The estimation procedure goes step by step from the total sample to the urban domain, to the urban literate subclass, and finally to urban literate women aged 20–29. Details are set out below.

Total sample (1st line)
Column (5) computed from equation 4.8: $\text{deft}' = \text{deft}/L^{1/2} = \frac{1.69}{1.21} = 1.40$

Column (4) computed from column (5): $\text{deft}'^2 - 1 = 1.40^2 - 1 = 0.96$

Column (8) computed from equation 4.9: $\text{se} = \frac{2.97}{\sqrt{5000}} \times 1.21 \times 1.40 = 0.071$

Urban domain (2nd line)
Column (4) computed from equation 4.18: $\text{deft}_s'^2 - 1 = 0.68 \times 0.96 = 0.65$

Urban literate subclass (3rd line)
Column (3): loss factor due to weighting, assumed same as whole urban domain, 1.09
Column (4) computed from equation 4.16: $\text{deft}_s'^2 - 1 = 0.65 \times 0.5^{1/2} = 0.46$

Urban literate aged 20–29 subclass (4th line)
Column (3): assumed same as whole domain, 1.09
Column (4) computed from equation 4.10: $\text{deft}_s'^2 - 1 = 0.46 \times 0.4 = 0.18$
Column (5) computed from column (4): $\text{deft}' = (0.18 + 1)^{1/2} = 1.09$
Column (6) computed from equation 4.8: $\text{deft} = L^{1/2} \cdot \text{deft}' = 1.09 \times 1.09 = 1.19$
Column (7) computed from equation 4.13 using the value of s for the total sample:

$$s = \left(\frac{3.07}{4.02}\right)^{1/2} \times 2.97 = 2.60$$

Column (8) computed from equation 4.9: $\text{se} = \frac{2.60}{\sqrt{400}} \times 1.09 \times 1.09 = 0.154$

Table 8 Illustration of extrapolation procedure for estimating sampling errors for subclasses

Subclass	m	n	$L^{1/2}$	$(\text{deft}'^2 - 1)$	deft'	deft	s	se	Assumed values of parameters; equations used
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Total sample	**4.02**	**5000**	**1.21**	0.96	1.40	**1.69**	**2.97**	0.071	se from eq. 4.9
Urban	**3.68**	**2000**	**1.09**	0.65	1.29	1.41	2.84	0.092	$c_s = 0.68$; eq. 4.18
Urban literate	**3.82**	**1000**	1.09	0.46	1.21	1.32	2.90	0.123	$M_s = 0.5$, $\alpha = 0.5$; eq. 4.16
Urban, literate and aged 20–29	**3.07**	**400**	1.09	0.18	1.09	1.19	2.60	0.154	$M_s = 0.4$; eq. 4.10

NOTE: Figures given are in bold type.

4.4 SAMPLING ERRORS FOR SUBCLASS DIFFERENCES

The variance of the difference between two means y_a and y_b is

$$\text{var}(\bar{y}_a - \bar{y}_b) = \text{var}(\bar{y}_a) + \text{var}(\bar{y}_b) - 2\,\text{cov}(\bar{y}_a, \bar{y}_b) \tag{4.19}$$

Except for the case when the two subclasses (a and b) come entirely from different PSUs, the covariance is generally positive, so that one can expect the inequality

$$\left(\frac{s_a^2}{n_a} \cdot L_a + \frac{s_b^2}{n_b} \cdot L_b\right) < \text{var}(\bar{y}_a - \bar{y}_b) < [\text{var}(\bar{y}_a) + \text{var}(\bar{y}_b)] \tag{4.20}$$

where the first expression on the left is the variance of $(\bar{y}_a - \bar{y}_b)$ for a random sample, and the expression on the right is that variance for the clustered sample disregarding the covariance term. In other words, the variance of the difference of two means from clustered samples shows the design effect of a positive intra-class correlation ($\text{deft}_d > 1$), but the effect is less than that for the separate means. This has been empirically demonstrated in many computations (eg Kish and Frankel 1974).

In a form similar to equation 4.9 we write

$$\text{var}(\bar{y}_a - \bar{y}_b) = \left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right) \cdot L_d \cdot \text{deft}_d'^2 \tag{4.21}$$

As was shown in table 1, the effect of weighting (L_d) tends to persist for differences between subclasses. Column (7) of table 5 shows that even for fairly large subclasses (most subclasses shown in the table are of size 1000–2000) the design factor, deft_d, for differences between subclasses is small, especially for the well-distributed demographic cross-classes; the values are somewhat larger for the less well-distributed socio-economic subclasses. (Note that column (7), headed 'deft_d', in table 5 shows design factors unadjusted for weighting. For Mexico and Thailand, the samples are self-weighting ($L_d = 1$), and it makes no difference. For Bangladesh, $L_d^{1/2} = 1.05$ and Indonesia $L_d^{1/2} = 1.18$. The deft_d shown in column (7) may be divided by the respective $L_d^{1/2}$ to obtain deft_d' as defined in equation 4.21. The average deft_d' for all variables for the demographic subclasses is 1.00 for Bangladesh and 1.03 for Indonesia. For socio-economic subclasses, the values are higher.)

The effect of the covariance term may be examined in terms of the following model based on equation 4.20:

$$\text{var}(\bar{y}_a - \bar{y}_b) = \beta^2\,[\text{var}(\bar{y}_a) + \text{var}(\bar{y}_b)] \tag{4.22}$$

where generally $0 < \beta < 1$; $\beta = 1$ when no covariance is present. Column (9) of table 5 illustrates the β values estimated on the basis of a large number of computations from WFS surveys. Generally β is in the range 0.9–1.0 (ie β^2 0.8–1.0); β values tend to be smaller for the well-distributed demographic subclasses, and also for groups of variables with larger deft.

An alternative, but particularly convenient form is as follows. Assuming $s_a^2 = s_b^2$ ($= s^2$, say) and that the weighting factor (L) is the same for the various subclasses and differences involved, we can write equation 4.20, using equation 4.10 with $M_s = n_a/n$ or n_b/n, as follows:

$$\frac{s^2}{n_d} \cdot L_d < \text{var}(\bar{y}_a - \bar{y}_b) < \frac{s^2}{n_d} \cdot L_d \left(1 + 2 \cdot \frac{n_d}{n} \cdot (\text{deft}_t'^2 - 1)\right) \tag{4.23}$$

where n_d is *half* the harmonic mean of subclass sizes n_a and n_b, ie

$$\frac{1}{n_d} = \frac{1}{n_a} + \frac{1}{n_b} \quad \text{or} \quad n_d = \frac{n_a \cdot n_b}{n_a + n_b}.$$

Even for fairly large subclasses, the range between the lower and upper limits in equation 4.23 tends to be small. For example, if $\text{deft}_t'^2 = 2$, $L_d = 1$ and $n_a = n_b = 0.2n$, we have $n_d = 0.1n$ and equation 4.23 becomes

$$10\frac{s^2}{n} < \text{var}(\bar{y}_a - \bar{y}_b) < 12\frac{s^2}{n}$$

If one takes $\text{var}(\bar{y}_a - \bar{y}_b)$ to be in the middle of this usually small range, then

$$\text{var}(\bar{y}_a - \bar{y}_b) \doteq \frac{s^2}{n_d} \cdot L_d \left(1 + \frac{n_d}{n}(\text{deft}_t'^2 - 1)\right), \tag{4.24}$$

which is identical to the relation developed earlier, equations 4.9 and 4.10, with n_d as the effective 'subclass size'. In other words, the variance of the difference between two subclasses is close to the variance for a subclass of size equal to half the harmonic mean of the two subclass sizes. If, for a particular variable and type of subclasses, the standard error can be reasonably approximated as a simple function of subclass size (as, for example, is implied by equations 4.9 and 4.10), then the *same* functional relationship (or tabulated values) may be used for subclass differences, with n_d, as defined above, taken as the subclass size.

4.5 DESIGN FACTORS FOR COMPLEX STATISTICS

Subclass differences represent a basic measure of *relation* between variables. Empirical findings about them lead to conjectures about design factors 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 analytic statistic, say γ, such as a correlation or regression coefficient:

1 $\text{deft}(\gamma) > 1$. In general, design factors 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 $\text{deft}(\gamma) < \text{deft}(\bar{y})$. Design factors 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 $\text{deft}(\gamma)$ is related to $\text{deft}(\bar{y})$. For variables with high $\text{deft}(\bar{y})$, values of $\text{deft}(\gamma)$ also tend to be high.

4 $\text{deft}(\gamma)$ tends to resemble $\text{deft}(\bar{y}_a - \bar{y}_b)$, the design factor for differences between means.

5 $\text{deft}(\gamma)$ tends to have measurable regularities for different statistics.

Based on the above, the authors propose a simple model

$$\text{deft}^2(\gamma) = 1 + k(\text{deft}^2(\bar{y}) - 1) \tag{4.25}$$

with $\text{deft}(\bar{y}) > 1$; and $k(0 < k < 1)$ being specific to a particular variable, type of statistic, and sample or sample subclass.

4.6 EXTRAPOLATION ACROSS VARIABLES

The discussion so far has considered the relationship of deft_s (for a subclass mean, difference, or other statistic) to deft_t for the total sample for a *given* variable. Generally, the relationships considered have been of the form

$$\text{deft}_s^2 = 1 + k_s(\text{deft}_t^2 - 1) \tag{4.26}$$

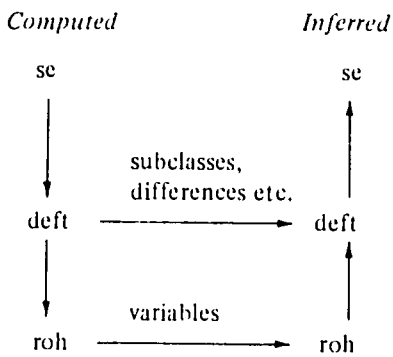
where k_s is a constant or some function of subclass size and type. Consequently, the inferential path was from standard error (se_t) computed for the whole sample, to corresponding deft_t, to deft_s and finally to se_s. Given the definition of roh, equation 4.5, the above equation implies an assumption about the relationship of roh_t and roh_s, such as

$$\frac{\text{deft}_s^2 - 1}{\text{deft}_t^2 - 1} = \frac{\text{roh}_s(\bar{b}_s - 1)}{\text{roh}_t(\bar{b}_t - 1)} = k_s$$

or

$$\frac{\text{roh}_s}{\text{roh}_t} = g_s, \text{ say,} \tag{4.27}$$

even though roh has not been introduced explicitly into the models considered. To infer sampling errors from one variable to another, it is necessary to speculate explicitly on the likely roh values. The inferential path may be shown schematically as follows:



The basic assumption is that roh depends upon the nature of the variable, so that the relative values of roh for two variables tend to persist as we move from the total sample to diverse subclasses and subclass differences, and possibly to other sample designs. Computed roh values are often unstable and such relationships hold only approximately. However, regular patterns have been found in suitably averaged values. Kish *et al* (1976) found some correspondence between the type of variable and its ranking according to roh values. Verma *et al* (1980) found the ranking of median rohs for groups of substantively similar variables to be consistent across a number of WFS surveys; furthermore, the ranking of *surveys* in different countries by median roh was found to be consistent across different groups of variables.

4.7 SAMPLING ERRORS FOR FERTILITY RATES

Computation of fertility rates from retrospective birth histories, for the total sample as well as for major domains, is an important objective of WFS surveys. Generally fertility rates are ratio estimates of the form

$$\text{rate} = \Sigma\,\text{birth}/\Sigma\,\text{exposure} \tag{4.28}$$

where the numerator is the accumulation (weighted if applicable) of births to a specified category of women during a specified time period, and the denominator is the accumulated length of 'exposure' to childbearing for the same category of women during the same period.

With different definitions of exposure, time periods and categories of women, various types of fertility rates can be defined (a detailed account is given in Verma 1980). To illustrate some issues relevant to sampling error, we will consider the conventional age-specific fertility rates (ASFRs). Here the births in the numerator are classified according to the period of occurrence (the one-year period prior to the interview, or specified calendar years, etc) and the age of the mother at birth of the child (20–24, 25–29, etc). The denominator is the number of person-years lived by women (irrespective of marital status) during the specified period and at the specified age.

In WFS surveys, two types of sample have been used for the detailed fertility interview: (a) a sample of all women within given age limits, irrespective of marital status, and (b) a sample of women within given age limits confined to women who are or have been married. In the computation of age-specific rates for the second case, it is necessary to adjust 'exposure' (equation 4.28) to include the never-married women who have been excluded from the sample interviewed. This can be done (assuming no births to never-married women) on the basis of information on the proportion ever-married from the *household* interview, which precedes, and forms the basis of selection for, the women's interview:

$$\text{rate} = \underbrace{\left(\frac{\Sigma\,\text{birth}}{\Sigma\,\text{exposure}}\right)}_{\text{ever-married women}} \times \underbrace{\left(\text{proportion ever married}\right)}_{\text{from household interview}}$$

$$\text{or} \quad r = r' \cdot p, \text{ say.} \tag{4.29}$$

r is a product of two ratios and its variance is, approximately,

$$\text{var}(r) = p^2\,\text{var}(r') + r'^2\,\text{var}(p) + 2p \cdot r'\,\text{cov}(p, r') \tag{4.30}$$

The covariance term can arise because r' is based on the same sample as p (or on a subsample of it). However, using 'all-women' samples from WFS surveys in Colombia and Kenya, on the basis of which var(r) as well as var(r') and var(p) can be directly calculated, Little (1982) found the covariance term not to be significant, so that from equations 4.29 and 4.30,

$$\frac{\text{var}(r)}{r^2} \doteq \frac{\text{var}(r')}{r'^2} + \frac{\text{var}(p)}{p^2}$$

Consider a fertility rate defined for a short reference period, say one year. For sampling error purposes, the chance of a woman having more than one birth during this period is negligible, so that the rate is equivalent to the *proportion* of women having a birth

during the one-year period. Hence, assuming a self-weighting sample, the variance of r may be written as:

$$se^2(r) = \left(\frac{r(1-r)}{e_1}\right) \cdot deft_1^2 \tag{4.31}$$

where suffix 1 indicates a reference period of one year and e_1 is the number of person-years lived by the sample of women during the one-year reference period, which is approximately equal to n, the sample size. As the reference period is increased to, say, p years, Little (1982) suggests the following modification to equation 4.31:

$$se^2(r) = \left(b_p^2 \frac{r(1-r)}{e_p}\right) \cdot deft_p^2$$

The design factor $deft_p$ tends to increase with increasing p; e_p is the total number of person-years lived by the sample of women during the p years reference period, and approximately

$e_p = p \cdot n$, that is, p times the sample size.

b_p is the 'birth correlation factor' and represents departures from the binomial model ($s^2 = r(1-r)$) which are not attributable to departures of the sample from simple random sampling. The factor takes into account the correlation between births to an individual woman. Empirically, the factor may not differ greatly from unity; Little reports average values between 0.99 and 1.05 for p = 3 years.

5 Presentation of Sampling Errors in Survey Reports

5.1 MODES OF PRESENTATION

Even when suitable computer programs are available for extensive computation of sampling errors, their presentation in a suitable form remains a problem in large-scale, multipurpose 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.

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 information on sampling errors must 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 them.

3 The presentation should be in a form which facilitates and encourages the proper interpretation and use of the information. It is better to provide approximate information which is more likely to be applied than to provide exact information which is hard to use.

4 Above all, the mode of presentation and the degree of detail given should suit the specific needs of particular categories of users.

Several categories of users may be distinguished. The first is the *general reader*, 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 second category is the *substantive analyst* engaged in primary or secondary analysis and reporting of results. This type of user requires access to more detailed results, and would expect to find not only direct estimates of sampling error for all major statistics, but also a general indication of the magnitude of sampling error to be expected for *any* statistic which may be derived from the survey.

The third category is the *sampling statistician* concerned with evaluating the statistical efficiency of the design adopted in the survey, or with designing samples for future surveys. This type of user is interested in relating the magnitude and components of sampling error to features of the sample design.

Before considering the question of presentation to suit different types of users, it is useful to remark on the general strategy. Even when the information of sampling

errors is presented in a summary form, it is desirable that this summarization should be *based on extensive computations.*[6] In a multi-subject survey, it is desirable to compute sampling errors for many substantive variables of different types. A very wide range of values of standard errors and design factors is generally found for diverse variables within the same survey. As noted earlier, an empirical basis for averaging results across different variables is generally much less certain than the averaging or modelling of variation across different sample subclasses for a given variable. Also, it is inadequate to single out arbitrarily one or a few variables as 'critical' survey variables for sampling error computations. The range of variables selected for computation should parallel the important aims of the survey, of its analysts and of its users (Kish *et al* 1976: 21).

It is desirable to repeat the computations for major domains of the sample for which separate results may be required. This is specially important if (as often is the case) the sample design varies from one domain to another. Further, it is important to compute sampling errors not only for the entire sample or its major geographic domains, but also for a range of subclasses and subclass differences. To generalize on the patterns of variation across subclasses, it is necessary to cover subclasses of different types, distribution and size.

Hence the general strategy should be to compute sampling errors for all important variables for the total sample, for each sampling domain, and for at least a moderate number of subclasses and differences. The larger the design factors (deft) for the total sample, the more important it is to investigate their variation for subclasses of diverse types and sizes.

5.2 FOR THE GENERAL READER

For the general reader, 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. As noted earlier, 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 their 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 terms such as 'sampling error', 'standard error' and 'confidence interval', etc, as discussed in section 2 above. These concepts should be illustrated by numerical examples. Gonzalez *et al* (1975) provide examples of an introductory text which may be used for this purpose. Their paper discusses the presentation of sampling and other survey errors at length, with many illustrations.

For the general reader, 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:

1 as absolute values of the standard error (se);

2 as relative values, standard error divided by the mean ($se/\bar{y}$); or

[6] Of course, in a multi-round survey, or in a series of similar surveys, the stability of variance patterns may obviate the need for detailed computations for each survey round (see, for example, Zarkovich: 1979). Nevertheless, for the set of surveys as a whole, valid conclusions regarding the behaviour of sampling errors can be based only on detailed computations at some stage.

3 in the form of probability or confidence intervals.

The preference between absolute and relative se will depend upon the nature of the estimate. The same value of relative se may be applicable to a number of estimates, for example for aggregates that vary greatly in size or in unit of measurement. In such cases, it is economical, as well as more illuminating for the reader, to present relative se.

However, absolute values of 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 per cent and standard error se = 2 per cent, the relative error is 5 per cent, 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 common practice, and that used in WFS First Country Reports, is to use the 95 per cent confidence interval, ie

estimate ± 2 · (standard error)

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 se) has similar values for a number of estimates, so that a single averaged value may suffice. Similarly, approximate values would suffice when 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. Sometimes, a little more detailed information may be provided by indicating different ranges of values of se 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 most WFS 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, the following system has been used to indicate the range of sample size (rather than of the standard error 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 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. (In WFS reports, this is done only for the text or summary tables, which are always accompanied by the full set of detailed cross-tabulations from which they are derived.) 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 the fact that the estimate is 'small' rather than 'large'. Consider two groups of women with the estimated mean numbe of children ever born as 6.1 and 6.2, and with standard error of the difference as 0.1. The 95 per cent confidence interval of the difference is $(6.2 - 6.1) \pm 2(0.1)$, that is -0.1 to 0.3, so that the difference is not statistically significant at the 95 per cent level. However, the results are substantively meaningful in that the difference is small, whatever its sign or exact magnitude.

5.3 FOR THE SUBSTANTIVE ANALYST

The substantive analyst will generally require more detailed information. 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:

1 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.

2 A graphic or tabular presentation of approximate standard errors (or other measures of sampling error) for a number of variables as a function of subclass size.

3 Similar information for differences between subclasses.

It may be necessary to produce summaries like (2) and (3) separately for different types of subclasses or for different sampling domains. The objective is to summarize results from detailed computations, smooth out random variability in 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.

Table 9 provides an illustration from the Indonesia Fertility Survey. It shows the approximate variation of standard error by subclass size, for each of the important survey variables. The table provides a good approximation for cross-classes (such as age groups of women) and for those subclasses which are distributed over most sample clusters, even if not uniformly. The latter would cover most socio-economic subclasses.

Table 9 Approximate value of standard error, by variable and subclass size (n_s), Indonesia Fertility Survey, 1976

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
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
First marriage dissolved	0.080	0.060	0.045	0.030	0.025	0.020	0.017	0.014	0.013	0.011	0.008	0.007
Remarried	0.065	0.050	0.035	0.025	0.020	0.017	0.015	0.013	0.012	0.009	0.008	0.007
Exposed	0.080	0.060	0.040	0.030	0.022	0.018	0.016	0.013	0.012	0.010	0.008	0.007
Children ever born[a]	0.450	0.340	0.240	0.180	0.130	0.110	0.090	0.080	0.070	0.060	0.050	0.040
Births in first 5 years	0.160	0.120	0.090	0.060	0.050	0.040	0.035	0.030	0.025	0.020	0.017	0.015
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	0.23
Births in past 5 years	0.165	0.125	0.090	0.065	0.050	0.042	0.035	0.031	0.027	0.022	0.018	0.017
Closed birth interval[b]	4.35	3.25	2.35	1.75	1.35	1.10	0.95	0.80	0.65	0.60	0.50	0.45
Open birth interval[b]	8.35	5.90	4.70	3.50	2.70	2.25	1.90	1.65	1.45	1.20	1.05	0.92
Months breastfed	0.98	0.78	0.52	0.37	0.28	0.23	0.20	0.17	0.14	0.12	0.10	0.09
Pregnant	0.045	0.032	0.024	0.018	0.013	0.011	0.009	0.008	0.007	0.006	0.005	0.004
Wants no more children	0.080	0.060	0.045	0.035	0.026	0.022	0.020	0.016	0.015	0.013	0.010	0.009
Prefers boy	0.080	0.060	0.045	0.032	0.024	0.020	0.018	0.014	0.013	0.011	0.009	0.007
Last child unwanted[c]	0.060	0.045	0.035	0.025	0.019	0.016	0.013	0.011	0.010	0.008	0.007	0.006
Additional number wanted[d]	0.265	0.190	0.140	0.105	0.080	0.070	0.060	0.050	0.045	0.037	0.032	0.028
Desired family size	0.335	0.260	0.195	0.145	0.115	0.100	0.085	0.075	0.065	0.055	0.045	0.040
Knows modern method	0.075	0.060	0.045	0.030	0.025	0.020	0.018	0.015	0.014	0.012	0.010	0.009
Ever used pill	0.065	0.050	0.040	0.030	0.022	0.019	0.017	0.014	0.013	0.011	0.009	0.008
Ever used IUD	0.040	0.030	0.025	0.018	0.014	0.012	0.009	0.008	0.007	0.007	0.006	0.005
Used any method	0.080	0.060	0.045	0.035	0.025	0.022	0.020	0.016	0.015	0.013	0.010	0.009
Used modern method	0.080	0.060	0.045	0.035	0.025	0.022	0.020	0.016	0.015	0.013	0.010	0.009
Using folk method	0.025	0.020	0.014	0.010	0.008	0.006	0.005	0.004	0.003	0.003	0.002	0.002
Using any method	0.080	0.060	0.045	0.035	0.025	0.022	0.020	0.016	0.015	0.013	0.010	0.009
Contracepting and wanting no more children	0.080	0.060	0.045	0.035	0.025	0.022	0.020	0.016	0.015	0.013	0.010	0.009

[a]For subclasses with mean < 2.5, multiply shown value of se by 0.5.
[b]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.
[c]For subclasses with proportion < 0.1, multiply shown values of se by 0.5. [d]For subclasses with mean < 0.5, multiply shown values of se by 0.5.
Source: Central Bureau of Statistics (1978)

Table 10 For standard error (se_d) of the difference between two subclasses of size n_1 and n_2, the appropriate sample base (n_d) to be used in table 9

	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

Procedure

To estimate standard error for the difference in mean/proportion between two subclasses of unweighted sample size n_1 and n_2 ($n_1 \leq n_2$, say) proceed as follows:

Read column in table 10 nearest to n_1 and row nearest to n_2. The cell at the intersection of these gives the appropriate size n_d to be used, for the given variable, in table 9.

If only the weighted subclass sizes are given, first use table 11 to obtain the unweighted sizes n_1 and n_2.

Source: Central Bureau of Statistics (1978)

The various footnotes to table 9 give very approximate adjustments to be made if the value of certain substantive estimates (means and proportions) for the subclass differs greatly from the value for the sample as a whole. Similar tables may be constructed for each geographic domain of the sample separately.

In fact, table 9 also provides approximate values of the standard error for subclass differences, using the approximation explained in section 4.4, that is, taking the effective sample size for a difference of two subclasses as half the harmonic mean of the two subclass sizes. Table 10 is used to determine the effective sample size n_d for the difference of two subclasses of size n_1 and n_2. Then this n_d is used in table 9 to estimate the approximate standard error for the difference.

For such information to be useful, the user must have access to sample sizes of individual cells in the detailed cross-tabulations. This generally presents no special problem in self-weighting samples. However, for weighted samples it is often not convenient to show both weighted and unweighted frequencies. Exact weighted frequencies are required to permit amalgamation of categories in the table, while unweighted frequencies are required only approximately, as an indication of the sampling error. When only one set of frequencies can be shown, it is preferable to show the weighted frequencies. Tables showing the approximate correspondence between weighted and unweighted frequencies for all major subclass categories and domains may then be required. An example is shown in table 11.

5.4 FOR 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:

Table 11 Factor by which weighted frequencies should be multiplied to obtain the corresponding unweighted sample size for various subclasses of the sample, by province and type of place of residence

Subclass	All Jawa-Bali	Type of place		Province[a]				
		Urban	Rural	Jawa Barat	Jawa Tengah	Yogyakarta	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	4.59

[a]Factor for Jakarta for all subclasses = 2.76.
NOTE: '–' Means not tabulated.
Source: Central Bureau of Statistics (1978)

1 Detailed information on standard errors and their pattern of variation with subclass type and size, as described in the previous section.

2 Similar information on design factors.

3 Information on roh values to permit extrapolation across variables and across designs.

4 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.

5 More generally, information on components of the sampling error for multi-stage designs.

As noted in the introduction, sample design is severely constrained by numerous practical considerations. Statistical efficiency is just one of the factors involved. Nevertheless, it is an important factor in making choices within the class of designs permitted, however narrowly, by considerations of practicality, quality control and cost.

References

Central Bureau of Statistics, Indonesia (1978). *Indonesia Fertility Survey 1976: Principal Report* (2 vols). Jakarta.

Gonzalez, M., J. Ogus, G. Shapiro and B. Tepping (1975). Standards for Discussion and Presentation of Errors in Survey and Census Data. *J. American Statistical Association 70 (351)*, pt II.

Hacettepe Institute of Population Studies (1980). *Turkish Fertility Survey 1978: First Report* (2 vols). Ankara.

Hansen, M.H., W.N. Hurwitz and M.A. Bershad (1961). Measurement Errors in Censuses and Surveys. *Bull. International Statistical Institute 38 (2)*: 359–74.

Kalton, G. (1977). Practical Methods for Estimating Survey Sampling Errors. *Bull. International Statistical Institute 46 (3)*.

Kish, L. (1959). Some Research Problems in Statistical Design. *American Sociological Review 24 (3)*.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Kish, L. and M.R. Frankel (1974). Inference from Complex Samples. *J. Royal Statistical Society B, 36*: 1–37.

Kish, L., R.M. Groves and K.P. Krotki (1976). Sampling Errors for Fertility Surveys. *WFS Occasional Papers* no 17.

Kish, L. and I. Hess (1959). On Variances of Ratios and their Differences in Multi-Stage Samples. *J. American Statistical Association 54*: 416–46.

Little, R.J.A. (1978). Generalized Linear Models for Cross-Classified Data from the WFS. *WFS Technical Bulletins* no 5.

Little, R.J.A. (1982). Sampling Errors of Fertility Rates from the WFS. *WFS Technical Bulletins* no 10.

Mahalanobis, P.C. (1944). On Large-Scale Sample Surveys. *Phil. Trans. Roy. Soc. B, 231*: 329–451.

O'Muircheartaigh, C.A. (1982). Methodology of the Response Errors Project. *WFS Scientific Reports* no 28.

O'Muircheartaigh, C.A. and A.M. Markwardt (1981). An Assessment of the Reliability of WFS Data. *World Fertility Survey Conference 1980: Record of Proceedings*, vol 3: 313. Voorburg, Netherlands: International Statistical Institute.

Tepping, B.J. (1968). Variance Estimation in Complex Surveys. *Proceedings Social Statistics Section, American Statistical Association*: 11–18.

United Nations (1982). *Non-Sampling Errors in Household Surveys: Sources, Assessment and Control*. Statistical Office, National Household Survey Capability Programme (DP/UN/INT-81-041/2).

Verma, V. (1978). CLUSTERS: a Package Program for the Computation of Sampling Errors. United Nations Economic Commission for Europe Conference of European Statisticians, meeting on problems relating to household surveys, Geneva.

Verma, V. (1980). Basic Fertility Measures from Retrospective Birth Histories. *WFS Technical Bulletins* no 4.

Verma, V. (1981a). Assessment of Errors in Household Surveys. *Bull. International Statistical Institute 49.*

Verma, V. (1981b). Sampling for National Fertility Surveys. *World Fertility Survey Conference 1980: Record of Proceedings,* vol 3: 389.

Verma, V. and M.C. Pearce (1978). Users' Manual for CLUSTERS. WFS Technical Paper no 770.

Verma, V., C. Scott and C. O'Muircheartaigh (1980). Sample Designs and Sampling Errors for the World Fertility Survey. *J. Royal Statistical Society A, 143,* pt 4: 431–73.

Woodruff, R. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *J. American Statistical Association 67*: 411–14.

Woodruff, R. and B. Causey (1976). Computerised Method for Approximating the Variance of a Complicated Estimate. *J. American Statistical Association 71*: 315–21.

World Fertility Survey (1975). Manual on Sample Design. *WFS Basic Documentation* no 3.

World Fertility Survey (1977). Guidelines for Country Report No. 1. *WFS Basic Documentation* no 8.

World Fertility Survey (1980). Data Processing Guidelines. *WFS Basic Documentation* no 10.

Yugoslavia Federal Statistical Office. Some Thoughts Concerning Sampling Errors. UN Economic and Social Council. Statistical Comm. and Economic Commission for Europe. Memo CES/AC. 48/8, 9 February 1978.

Zarkovich, S.S. (1979). Stability of Variance Patterns. *J. Indian Society of Agricultural Statistics 31 (1)*: 23–48.

Appendix A- A Brief Description of the CLUSTERS Package

1 INTRODUCTION

One of the main obstacles in computing and presenting sampling errors along with substantive survey results has been the non-availability of an adequate, efficient and easy-to-use computer software package for sampling error calculations. To rectify this situation, WFS developed, and distributed for a nominal charge, a fairly modest, simple and fast package called CLUSTERS.[7] The package has facilitated the inclusion of sampling error results in most WFS First Country Reports. However, the usefulness of CLUSTERS is by no means limited to WFS-type surveys. The following description of the main features of the package is provided to encourage agencies engaged in sample surveys to utilize the package, and undertake routine and extensive computation of sampling errors.[8]

2 BASIC REQUIREMENTS FOR A SOFTWARE PACKAGE

With CLUSTERS, an attempt has been made to meet the basic requirements for a general and widely usable software package for calculation of sampling errors for descriptive sample surveys. These requirements can be outlined as follows:

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 extent of sampling errors. However, the program should not be limited to a particular sample design; it should not assume particular models like 'paired selection of primary sampling units' in order to estimate variances.

5 It should be able to handle weighted data.

6 As far as possible, the program should not require any particular arrangement of 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 computed values may assist users to extrapolate

[7] Computation and Listing of Useful STatistics on ERrors of Sampling.
[8] The following description is adapted from Verma (1978). Further details are available in the Users' Manual (Verma and Pearce 1978).

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.

3 MAIN FEATURES OF CLUSTERS

CLUSTERS is a FORTRAN IV based software package (it also requires a standard sort program). It uses approximately 50 K bytes of core storage though if more is available for work areas, more calculations can be done in one run of the program. However, 50 K bytes is enough for an average number of variables and subclasses. The package is not machine dependent, and has been installed on a variety of machines (IBM 360/370, ICL 2900, HP 3000, CDC 6000).

Below the main features of CLUSTERS are summarized in relation to the basic requirements discussed above.

Handling of Different Variables and Sample Subclasses

We note that subclasses for sampling error calculations usually are defined in terms of the characteristic used in the cross-classification of the substantive results from the surveys. Often the same system of cross-classification is relevant to all (or most) survey variables. Variables like family size or prevalence of contraceptive use may all be presented after classification of the sample by characteristics such as age or socio-economic background of the units of analysis.

Making use of these common features, the calculations to be performed are specified in CLUSTERS in terms of a rectangular 'variable by subclass' matrix. Sampling errors are then computed for all variables over each subclass (and automatically over the whole sample) in the specified set. In addition, CLUSTERS automatically computes sampling errors for each subclass, treating it as a characteristic distributed over the entire sample. As an example, if sampling errors for 20 variables over 15 sample subclasses are to be computed (a typical WFS survey requirement), it is not necessary to specify $20 \times 15 = 300$ 'problems' separately, but only $20 + 15 = 35$ variables and subclasses.

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.

Separate Results for Geographical Regions

The entire set of calculations for variables over sample subclasses and for differences between subclass pairs can be repeated for the separate geographical regions into 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.

Sample Structure

CLUSTERS computes sampling errors taking into account the actual sample design, in particular the clustering and stratification of the sample. The basic units involved in the computational formulae are the primary sampling units (PSUs), ie the first or highest stage units selected into the sample. The procedure is roughly as follows: for any variable under study, a summation (weighted if applicable) is made over the values of the variable for all individual cases (belonging to a particular sample subclass) in each PSU. The PSU totals are then differenced from a mean of all sample PSUs within each stratum according to formulae described in the Users' Manual. These differences are then squared and pooled over the whole sample (or over each geographical region, if applicable) and divided by an appropriate constant to produce estimates of sampling variance.

For a multi-stage sample, the procedure does not split the overall variance into separate components associated with the individual stages. Hence all that is required regarding specification of the sample structure is an identification of the PSU, stratum and geographic region (if applicable) for each individual case (ie each ultimate sampling unit). One of the noteworthy features of CLUSTERS is the fair degree of flexibility regarding the form of this identification; restructuring or recoding of the input data is not normally required.

Weighted Data

CLUSTERS handles non-self-weighting samples, ie samples in which the ultimate units need to be weighted to compensate for differences in probabilities of selection or for defects in sample implementation, eg non-response. These sample weights may be scaled arbitrarily and specified either as a data field on each individual record or simply in terms of the identification code for each of the 'higher stage' units mentioned in the previous paragraph.

Recoding of Input Data

It is often necessary to recode raw input data before the required statistics like proportions, means, or ratios and their standard errors can be calculated. For this purpose, CLUSTERS includes a limited set of recoding facilities. These can define new variables on the basis of one or more input data fields. Though using these facilities is not always the most economical means of recoding variables, they are simple to use and have been found quite versatile.

Derived Statistics

In addition to standard errors, CLUSTERS outputs two derived statistics, namely design factor (deff) and rate of homogeneity (roh). They provide the basis for generalizing the computed results to other variables and subclasses of the particular sample, and possibly also to other sample designs.

4 LIMITATIONS

The main limitations of the package are as follows.

1 It cannot handle hierarchical data files. The file must be rectangular with no non-numeric codes.

2 Analysis of variance into components attributable to different sampling stages is not included. However, by repeating the calculation, ignoring one or more higher stages, the effect of those stages can be isolated approximately. For illustrations of this approach see Verma *et al* (1980).

3 CLUSTERS is confined to descriptive statistics, such as proportions, percentages, means and ratios. Differences of ratios of only a specific (but by far the most commonly encountered) type are handled. It is assumed that the two ratios being differenced are defined by the *same pair* (numerator and denominator) of variables, but over different subclasses of the sample; the subclasses may overlap and need not be exhaustive. The package does not handle more complex statistics such as general linear combinations of ratios, products or ratios of ratios, nor of course, regression and correlation coefficients, etc.