Model Quality Report in Business Statistics

Mats Bergdahl, Ole Black, Russell Bowater,
Ray Chambers, Pam Davies, David Draper, Eva Elvers,
Susan Full, David Holmes, Pär Lundqvist,
Sixten Lundström, Lennart Nordberg, John Perry,
Mark Pont, Mike Prestwood, Ian Richardson,
Chris Skinner, Paul Smith, Ceri Underwood, Mark Williams

General Editors: Pam Davies, Paul Smith

Volume I

Theory and Methods for Quality Evaluation

Preface

The *Model Quality Report in Business Statistics* project was set up to develop a detailed description of the methods for assessing the quality of surveys, with particular application in the context of business surveys, and then to apply these methods in some example surveys to evaluate their quality. The work was specified and initiated by Eurostat following on from the Working Group on Quality of Business Statsitics. It was funded by Eurostat under SUP-COM 1997, lot 6, and has been undertaken by a consortium of the UK Office for National Statistics, Statistics Sweden, the University of Southampton and the University of Bath, with the Office for National Statistics managing the contract.

The report is divided into four volumes, of which this is the first. This volume deals with the theory and methods for assessing quality in business surveys in nine chapters following the survey process through its various stages in order. These fall into three parts, one dealing with sampling errors, one with a variety of non-sampling errors, and one covering coherence and comparability of statistics.

Other volumes of the report contain:

- a comparison of the software methods and packages available for variance estimation in sample surveys (volume II);
- example assessments of quality for an annual and a monthly business survey from Sweden and the UK (volume III);
- guidelines for and experiences of implementing the methods (volume IV).

An outline of the chapters in the report is given on the following page.

Acknowledgements

Apart from the authors, several other people have made large contributions without which this report would not have reached its current form. In particular we would like to mention Tim Jones, Anita Ullberg, Jeff Evans, Trevor Fenton, Jonathan Gough, Dan Hedlin, Sue Hibbitt and Steve James, and we would also like to thank all the other people who have been so helpful and understanding while our attention has been focussed on this project!

Outline of Model Quality Report Volumes

Volume I

Volume II

Volume III

Volume IV

Contents

Part 1: Sampling Errors

Part 2: Non-sampling errors

Part 3: Other Aspects of Quality

Part 4: Conclusions and References

1 Methodology overview and introduction

Paul Smith, Office for National Statistics

1.1 General structure

This volume covers the theory and methods for assessing quality in business surveys under eight main headings. The main body of the report is divided into nine chapters, with the probability sampling main heading split into two chapters. The non-sampling error sections follow the classification of the Eurostat working group on Quality of Business Statistics. The chapters are

2. Probability sampling: basic methods
3. Probability sampling: extensions
4. Sampling errors under non-probability sampling
5. Frame errors
6. Measurement errors
7. Processing errors
8. Nonresponse errors
9. Model assumption errors
10. Comparability and coherence

These fall into three parts, with chapters 2-4 dealing with sampling errors (part 1), chapters 5-9 with various aspects of non-sampling errors (part 2) and chapter 10 forming a part on its own (part 3). The coverage of each chapter is described in summary in section 1.2, and the ideas are synthesised and linked to the Model Quality Reports in the final chapter, chapter 11. References to other work mentioned in this volume appear at the end, and the notation generally follows Särndal, Swensson & Wretman (1992) except where further notation is required, in which case it is defined.

1.2 A guide to the contents

1.2.1 Total survey error

It is sensible to try to link the methods in these sampling and non-sampling error chapters into a common framework (a) as a guide to what is of most interest and relevance and which source of error is likely to be most important in a given context, and (b) to help in navigation through the topics contained in the various chapters. This is especially important in some of the non-sampling error chapters where topics will often fit comfortably under more than one heading, and it may not be immediately obvious where to look for information on a particular topic.

The best concept for providing a unifying framework is the concept of **total survey error** (Groves 1989), which embodies the difference between the survey estimate and the conceptual "real" or "true" value. In business surveys the real value (total sales by manufacturing industries, for example) mostly has a foundation in reality – if it were possible to look at every manufacturing business' sales and record them accurately, we could arrive at the real value. For other statistics such as the average price movement the “true value” is not well-defined and this construct breaks down. So, assuming that the real value is well-defined, we can imagine that we want to measure the difference between our survey estimate and the true value.

Consider the problem of estimating a total $\sum_U y_i$ of a variable *y* across a population *U*. The typical estimator takes the form $\sum_s w_i y_i'$, where w_i is the survey weight, y_i' is the reported value of y_i and the sum is over the sample *s*. The total survey error is then

$$\text{total survey error} = \sum_s w_i y_i' - \sum_U y_i$$

and this may be broken down into two components (see Groves, 1989, p.11):

$$\begin{aligned}\text{error from observation} &= \sum_s w_i y_i' - \sum_s w_i y_i = \sum_s w_i (y_i' - y_i)\\ \text{error from non - observation} &= \sum_s w_i y_i - \sum_U y_i\end{aligned}$$

The first (observation error) component reflects measurement errors, as well as processing, coding and imputation errors and would disappear if the recorded values y_i' were equal to the true values y_i. The second (non-observation error) component reflects sampling errors, frame errors and nonresponse errors and would disappear if the units *s* upon which the estimate is based comprised precisely the target population *U*.

The total survey error provides an overall measure of quality. The problem is how to assess its magnitude. To measure the sampling error it is usual to set up a model for the distribution of the sampling error and then to estimate the characteristics of this distribution. Usually, it is assumed (the assumption being based on asymptotic theory) that this sampling distribution is approximately normal and centred at zero so that the only task is to estimate the variance of the distribution. To extend this idea to total survey error it is necessary to set up a model for the distribution of the other components of error.

Total survey error can be considered in a different way too – broken down into two components, a difference which is approximately invariant over repetitions of the survey, the **bias**, and a difference which varies with different repetitions of the survey, the **variance**. The repetitions used in this definition are often hypothetical, that is the survey is not actually repeated. We explore these two types of error in more detail below.

The bias and variance together contribute to a measure of the total survey error, called the **mean squared error** (mse), such that

$$\text{mse} = \text{bias}^2 + \text{variance}$$

also sometimes expressed as its square root, the root mean square error (rmse). Both the bias and the variance are made up of several component terms corresponding to particular types of errors. In the case of the bias some of these components will almost certainly cancel each other out (we say that there are positive and negative biases), so that the overall bias will be the net of these effects. Variances are always non-negative[1] and so will cumulate over components. If all the relevant biases and variances are included in calculating the mse, it will be a good estimator of the total survey error.

This gives us two broad approaches to many errors. We can treat the response of a given unit as fixed for any occasion when it is included in the sample (a kind of deterministic approach). That is, if a business is included in the sample, we assume that it always makes the same response/nonresponse decision, always gives the same answers on the questionnaire, and so on. This almost always leads us to estimate biases. Alternatively we can consider that a business's response/nonresponse decision arises from some probability distribution, and that its answers also come from some distribution, in which case most of the errors will additionally have a variance component. This latter approach is akin to the model-based sampling approach (section 2.3.2), as we assume a *superpopulation* of possible outcomes with the sampling forming only one component of determining which outcomes we actually observe in the survey. We will use this distinction in approach between deterministic and superpopulation models in discussing the errors which make up total survey error.

1.2.2 Sampling errors

Certain assumptions and models are required to estimate the components of total survey error, and we begin by considering random sampling mechanisms; in this section we assume that all survey stages after sampling are error-free. When a survey is to be conducted, the sample can be selected according to some probability mechanism. At least conceptually we can select more then one sample using the *same* probability mechanism (by running the selection process several times), and each sample would result in a different estimate if the survey were actually run, simply because different units would be included in the sample. Each of these potential estimates would in general be different from the true total. We have here the situation that the survey estimates are different by repetition over different samples, and we can measure how much these estimates differ from their mean on average, using the average distance of the sample elements from their mean to estimate the average distance of population elements from the mean. This gives us a variance, the sampling variance. Over all possible different samples, the mean of the estimates is the same as the true value (still assuming no other errors); in practice we normally have only one sample, and have to use the mean of that sample to approximate the true population mean. Effectively , as mentioned in section 1.2.1, we assume that the sampling error is centred around the estimate we do have.

Chapter 2 covers the theory and methods which give rise to sampling error and sampling error estimates using firstly the design-based and model-assisted approaches, under which different models of the relationship between a survey response and known auxiliary values are used to improve the estimation. These approaches basically involve accounting for the selection probabilities from the sampling in all the estimation and variance calculation in an appropriate way. This chapter also introduces the model-based approach, which assumes that

[1] Unless estimated by a variance component model; if a negative variance is obtained it probably indicates that the model is inappropriate.

the survey responses are realisations from an hypothetical infinite population of possible outcomes. In this case, with an appropriate model the selection probabilities are *ignorable*, that is they have no effect on the estimation or variance estimation and do not need to be included explicitly.

Chapter 3 takes these two approaches and extends them from straightforward estimation methods to more complicated statistics, including estimation of changes, estimation for domains (subsets of the population) and estimation in the presence of outliers. There is also a summary of some work on the variability of a multisource indicator, which considers the effects of the variability of different series which go to make up an index on its total variance.

Consider now sample selection mechanisms which are not based on probability. In these cases the types of errors we obtain depend on the actual mechanism of selection. If repetition has no effect on the sample composition (that is, the same sample elements are chosen every time), then the difference between the survey estimate and the true value is constant over repetitions: it is a (pure) bias. If the sample can be different over repetitions, then there will be a range of potential estimates, and there will be a variance component and a bias. In practice the two effects may not be separately estimable, or even estimable at all if the true value is unknown (which is typically the case). This subject is addressed in chapter 4 (nonprobability sampling), concentrating particularly on cut-off sampling and voluntary sampling (samples obtained from voluntary surveys), but also mentioning quota sampling and judgemental sampling.

1.2.3 Non-sampling errors

Now relaxing the assumption from section 1.2.2 that everything else apart from sampling is perfect, let us consider the other possible errors. These are arranged to follow approximately the order of processing in a business survey.

Frame errors – contributing mainly to the bias component of the total error – are discussed in chapter 5. These errors generally stem from differences between frame- and target populations. Hence problems of under- and over-coverage are important. Since business populations usually change rapidly, the updating of units and of variables attached to these units become important. Delineation of businesses into different types of units (local units, kind-of-activity units etc) is another activity with a large impact on frame quality. All of these issues are dealt with in chapter 5.

Measurement errors are errors which are introduced when trying to get the desired information from contributors. In chapter 6, we look at a measurement error model for how answers vary over different (conceptual) repeated questionings, and this contributes to the variability of the estimates by giving a variable measurement for a single respondent. Measurement errors are likely to contribute to both components – bias and variance – of the total error but they are often difficult or expensive to assess, especially in cases where follow-up studies become necessary. Yet measurement errors may often have a large influence on accuracy in business surveys. Approaches to detection and assessment of measurement errors are discussed in chapter 6.

Processing errors are discussed in chapter 7. These are errors connected with data handling activities – whether manual or automated – such as data transmission, data capture, coding and data editing. A particular form of processing errors, called systems error in chapter 7, are errors arising from software and hardware. It is difficult to envisage a probability mechanism with a real interpretation for systems errors, and in fact they are very difficult to measure at all. Processing errors in general may contribute to both components – the bias and the variance – of the total error although the bias is likely to be the more important one.

Nonresponse, treated in chapter 8, arises when a sampled unit fails to provide complete responses to all questions asked in a survey. There are two ways of considering nonresponse in a fixed sample. The deterministic approach assumes a fixed but unknown response indicator value (1 if value is recorded, 0 if value is missing) for every unit in the sample. The stochastic approach treats the response indicator variables as outcomes of random variables. The nature of errors arising from nonresponse then depends on assumptions about this random mechanism. The stochastic approach is the one followed in chapter 8. Methods to measure or indicate the impact of nonresponse on accuracy are treated. This chapter also treats implications of nonresponse such as bias, variance inflation and effects of confusing nonresponse with over-coverage. Re-weighting and imputation methods to compensate for bias caused by nonresponse are discussed.

Chapter 9 discusses errors and inaccuracy caused by using model assumptions concentrating on estimation problems and types of models which are not mentioned elsewhere. The aim of introducing a model may be to reduce variance and/or to reduce bias, but there is also a risk of introducing bias if the model is not well chosen. Small area estimation is one part of the survey process where models are important, benchmarking another (note that calibration belongs to sampling errors; the idea is similar but the technique different). Non-ignorable nonresponse is discussed here, although it has strong links to the non-response methods in chapter 8. The discussion of cut-off sampling was started in chapter 4, non-probability sampling, and it is continued here, emphasising the use of models to estimate for the part of the population that was cut off. Another reason for using models is to help to compensate for a lack of up-to-date information, for example on weights in chained price indices, a problem which is introduced in this chapter. Seasonal adjustment is also described, including comments on the software in use; assessment of the resulting accuracy is a difficult matter.

1.2.4 Comparability and coherence

This is an area which does not fit under the usual definition of total survey error, because it does not deal with the errors in a single survey, but instead considers how well two or more sets of statistics can be used together. This chapter covers definitions in theory and in practice, accuracy, different co-ordination activities, and comparability of surveys over time and national boundaries. Both user and producer perspectives are considered, and illustrations are given.

1.2.5 Concluding remarks

The final chapter in this volume, chapter 11, links the concepts described in this introduction and draws out the important themes for assessing total survey error in some given contexts. It also corresponds with chapter 2 of the Implementation Guidelines (volume IV), which provides a summary of the methods described in this volume as they are applied in the Model Quality Reports.

There is an example running through the sampling error chapters (2 and 3), and which also appears in chapters 4, 8 and 9, which corresponds strongly with the Annual Business Inquiry in the UK, which is the annual structural survey example from the UK in the Model Quality Reports (volume III, chapter 3).

Part 1: Sampling Errors

2 Probability sampling: basic methods

Ray Chambers, University of Southampton

2.1 Basic concepts

Many scientific and social issues revolve around the distribution of some type of characteristic over a population of interest. Thus the number of unemployed people in a country's labour force and the average annual profit made by businesses in the private sector of a country's economy are two key indicators of that country's economic well-being. The first of these numbers depends on the distribution of labour force states among the individuals making up the country's labour force while the second is determined by the distribution of annual profits achieved by the country's businesses. Both these numbers are typically measured by sample surveys. That is, a sample of individuals belonging to the country's labour force is surveyed and their employment/unemployment statuses determined. Similarly a sample of private sector businesses is surveyed and their annual profits measured. In both cases the information obtained from the survey can be used to "infer" the unknown corresponding value (unemployment total or average profit) for the country.

2.1.1 Target population and sample population

Since in general it is meaningless to talk about a sample without referring to what it is a sample *of*, the concept of a *population* from which a sample is taken is basic to sample survey theory. In the examples above there are two populations – the population of individuals making up the labour force of the country, and the population of businesses making up the private sector economy of the country.

In general, however, the population from which a sample is taken, and the population of interest can and do differ. The *target population* of a survey is the population at which the survey is aimed, that is the population of interest. However, a target population is not necessarily a population that can be surveyed. The actual population from which the survey sample is drawn is called the *survey population*. A basic measure of the overall quality of a sample survey is the coverage of the survey population, or the degree to which target and sample population overlap. Assessment of this quality is considered in Chapter 5. Here we shall assume there is no difference between the target and survey populations. That is, we have *complete coverage*. From now on we will just refer to the population.

2.1.2 Sample frames and auxiliary information

A standard method of sampling is to select the sample from a list (or series of lists) which enumerate the units (individuals, businesses, etc) making up the sample population. This list is called the (*sample*) *frame* for the survey. Existence of a sample frame is necessary for the use of many sampling methods. Furthermore, application of these methods often requires that a frame contain more than just identifiers (for example, names and addresses) for the units making up a sample population. For example, stratified sampling requires the frame to contain enough identifying information about each population unit for its stratum membership to be determined. In general, we refer to this information as *auxiliary information*. Typically, this auxiliary information includes characteristics of the survey population that are related to the variables measured in the survey. These include stratum identifiers and measures of "size". For economic populations, the latter correspond to values for each unit in the population which characterise the level of economic activity by the unit.

The extent to which the sample frame enumerates the sample population is another key measure of sample survey quality. This issue is considered in Chapter 5. In what follows however we shall assume a sample frame exists and is perfect. That is, it lists every unit in the population once and only once, and there is a known number N of such units.

2.1.3 Probability sampling

A probability sampling method is one that uses a randomisation device to decide which units on the sample frame are in the sample. With this type of selection method, it is not possible to specify in advance precisely which units on the frame make up the sample. Consequently such samples are free of the (often hidden) biases that can occur with sampling methods that are not probability-based. In what follows we make the basic assumption that the probability sampling method used is such that every unit on the frame has a non-zero probability of selection into the sample. This assumption is necessary for validity of the design-based approach to survey estimation and inference described in section 2.3.1 below. Some relevant theory for the case where a non-probability sampling method is used is set out in Chapter 4.

2.2 Statistical foundation

As noted earlier, the basic aim of a sample survey is to allow inference about one or more characteristics of the population. Such characteristics are typically defined by the values of one or more population variables. A population variable is a quantity that is defined for every unit in the population, and is observable when that unit is included in the sample. In general, surveys are concerned with many population variables. However, most of the theory for sample surveys has been developed for the case of a small number of variables, typically one or two. In what follows we adopt the same simplification. Issues arising out of the need to measure many variables simultaneously in a sample survey are considered in section 2.3.4.

2.2.1 *Y* and *X* variables

Associated with each unit in the population is a set of values for the population variables. Some of these are recorded on the frame, and so are known for every unit in the population. We refer to these auxiliary variables as X-variables. The others constitute the variables of interest (the study variables) for the survey. These are not known. However we assume that their values are measured for the sampled units, or can be derived from sample data. We usually refer to these variables as Y-variables.

For example, the quarterly survey of capital expenditure (CAPEX) carried out by the U.K. Office for National Statistics (ONS) has several study (*Y*) variables, the most important being acquisitions, disposals and the difference between acquisitions and disposals, the net capital expenditure. The frame for this survey is derived from the Inter-Departmental Business Register (IDBR) of the ONS. There are a number of *X*-variables on the survey frame, the most important of which are the industry classification of a business (Standard Industry Classification), the number of employees of the business and the total VAT turnover of the business in the preceding year.

2.2.2 Finite population parameters

The population characteristics that are the focus of sample surveys are sometimes referred to as its targets of inference. In general, these targets are well-defined functions of the population values of *Y*-variables, typically referred to as parameters of the population. Any population covered by a frame-based survey is necessarily finite in terms of the number of units it contains. Such a parameter will be referred to as a *finite population parameter* (*FPP*) in what follows in order to distinguish it from the parameters that characterise the infinite populations used in standard statistical modelling. Some common examples of *FPP*'s are:

- the population total and average of a *Y*-variable;
- the ratio of the population averages of two *Y*-variables;
- the population variance of a *Y*-variable;
- the population median of a *Y*-variable.

2.2.3 Population models

A population of *Y*-values at any one point in time represents the outcome of many chance occurrences. However, this does not mean that these values are completely arbitrary. There is typically a structure inherent in a set of population values that can be characterised in terms of a *model*. Such models are usually based on past exposure to data from other populations very much like the one of interest, or subject matter knowledge about how the population values ought to be distributed. Consequently this model is not causal – it does not say how these *Y*-values came to be – but descriptive, in the sense that it is a mathematical description of their distribution. In many cases this model is itself defined in terms of parameters which "capture" these distributional characteristics.

A standard way of specifying such a statistical model is in terms of an underlying stochastic process. That is, the *N* values constituting the finite population of interest are assumed to be realisations of *N* random variables whose joint distribution is described by the model. If this approach is taken, then the model itself is referred to as a *superpopulation model* for the finite population of interest. The parameters that characterise this model are typically unknown, and are referred to as the superpopulation parameters for the population. Unlike *FPP*'s, superpopulation parameters are not real – they can never be known precisely, even if the superpopulation model is an accurate depiction of how the finite population values are distributed and every population value is known. Some examples of such superpopulation parameters are moments (means, variances, covariances) of the joint distribution of the *Y*-

variables defining the population values and related quantities (for example regression coefficients).

2.2.4 Sample error and sample error distribution

Once a sample has been selected, and sample values of *Y*-variables obtained, we are in a position to calculate the values of various quantities based on these data. These quantities are typically referred to as *statistics*. The aim of sample survey theory is to define two types of statistics:

(i) estimates of the *FPP*'s of interest;

(ii) quality measures for the estimates in (i).

In this report we will be mainly concerned with the second type of statistic above, that is statistics measuring the quality of the estimates. However, before we can describe how such statistics can be derived, we need to discuss the concepts of sample error and sample error distribution.

The sample error of a survey estimate is just the difference between its observed value and the unknown value of the *FPP* of which it is an estimate. Clearly one would expect a high quality survey estimate to have a small sample error. However, since the actual value of the *FPP* being estimated is unknown, the sample error of its estimate is also unknown. But this does not mean that there is nothing we can say about this error. The method by which the sample is chosen, and the superpopulation model for the population, allow us to specify a variety of distributions for the sample error. In turn, this allows us to use statistical methods to measure the quality of the survey estimate in terms of the characteristics of these distributions.

Before going on to describe how these distributions are derived and interpreted, it is important to note that this quality measurement relates to a quantity (the sample error) which assumes that there are no other sources of error in the survey. In reality, there are many other sources of error (frame error, nonresponse error, measurement error, model specification error, processing error) in a survey. Methods for assessing these are discussed in Part 2 of this report.

2.2.5 The repeated sampling distribution vs. the superpopulation distribution

There are two standard ways of defining a distribution for a sample error. One is its *repeated sampling distribution*. This is the distribution of possible values this error can take under repetition of the sampling method. Conceptually, this corresponds to repeating the sampling process, selecting sample after sample from the population, calculating the value of the estimate for each sample, generating a (potentially) different sample error each time and hence a distribution for these errors.

The other way of defining a distribution for a sample error is in terms of the superpopulation distribution. Under this distribution the sample estimate as well as the FPP are both based on realisations of the *Y*-variables that define the population values. Consequently the sample error is also a *random variable* with a distribution defined by the superpopulation model.

Operationally this distribution corresponds to the range of potential values the sample error can take given the range of potential values for the population *Y*-variables under this model.

There are fundamental differences between these distributions. The repeated sampling distribution treats the population values as fixed. Consequently the source of variability underlying this distribution is the sample selection method. Sample selection methods that are not probability based are therefore not suited to evaluation under this distribution. In contrast, the superpopulation distribution treats the sample as fixed. That is, the underlying variability in this case arises from the uncertainty about the distribution of *Y*-values for the sample units and non-sample units, but the sample/non-sample distinction is fixed according to that actually observed.

To distinguish between these two distributions, we use a subscript of p in what follows to denote expectations, variances, etc, taken with respect to the repeated sampling distribution, and a subscript of ξ to denote corresponding quantities taken with respect to the superpopulation distribution.

There are statistical arguments for and against the use of these two distributions for the sample error when we want to characterise the quality of the actual sample estimate. Basically, the repeated sampling (or randomisation) distribution of the sample error is viewed as appropriate for measuring the quality of a survey design, that is the method used to select the sample. This is because it reflects our uncertainty about which sample will be chosen prior to the actual choice of sample. However, both methods have been used to characterise uncertainty about the size of the sample error after the sample data are obtained. The argument for using the randomisation distribution involves the assumption that these data do nothing to change the source of our uncertainty, they just provide us with a means to measure it. We still characterise uncertainty by the distribution of sample errors associated with samples that might have been chosen but were not. In contrast, use of the superpopulation distribution essentially comes down to saying that the population *Y*-values, being unknown, represent the true source of uncertainty as far as survey inference is concerned. In particular, after the sample data are obtained we have no uncertainty about which sample was selected, but we still have uncertainty about the population *Y*-values defining the *FPP* of interest. In this report we will develop measures based on both distributions, indicating their strengths and weaknesses where appropriate.

2.2.6 Bias, variance and mean squared error

In order to use a distribution for the sample errors to measure the quality associated with the actual sample estimate, we need to specify the characteristics of this distribution that are appropriate for this purpose. Statistical practice essentially focuses on two such characteristics – the central location of the distribution, as defined by its mean or *expectation*, and the spread of this distribution around this mean, as defined by its *variance*. Often both are combined in the *mean squared error*, which is the variance plus the squared mean. The mean of the sample error distribution is typically referred to as the bias of the estimation method, so the mean squared error becomes variance plus squared bias.

A high quality estimate will be associated with a sample error distribution that has bias close to or equal to zero and low variance. In this case we can be sure that the observed value of the estimate will, with high probability, be close to the unknown *FPP* being estimated. Consequently we focus on the bias and variance of the sample error distribution as the key quality measures of a sample estimation method. In the next section we develop expressions for these quantities, together with relevant methods for estimating them from the sample data. In doing so we focus on one *FPP* that is of particular interest in many survey sampling situations. This is the *FPP* defined by the total t of the values taken by a single Y-variable.

2.3 Estimates related to population totals

Let U denote the finite population of interest, and let $j \in U$ denote the N units making up this population. For each unit we assume that a Y-variable is defined, with the realised (but unknown) value of this variable for the j^{th} unit denoted by y_j. The total of the N values of this Y-variable in the population will be denoted t. Following common practice we do not distinguish between y_j as a realisation (that is a number) and y_j as the random variable that led to that realisation. It will be clear from the context what particular interpretation should be placed on this quantity. Similarly, we will not distinguish between an estimate (a realised value) and an estimator (the procedure that led to the realised value).

2.3.1 The design-based approach

This approach, often referred to as *design-based theory*, evaluates an estimate of t in terms of the repeated sampling distribution of its sample error. That is, a good estimate for t is defined as one for which the associated sample error is known to be a "draw" from a repeated sampling distribution that has either zero bias or bias that is approximately zero and a small variance. As will become clear below, the usefulness of this approach depends on whether or not a random method with known sample inclusion probabilities is employed for sample selection.

2.3.1.1 Sample inclusion probabilities

In order to generate this repeated sampling distribution we need to introduce the concept of a *sample inclusion indicator*. This is a binary valued random variable that takes the value 1 if a unit is included in sample and is zero otherwise. We denote it by I in what follows. Clearly the distribution of I_j depends on the process used to choose the sample. Suppose now that this process is random in some way. Then we can put $\pi_j = \Pr_p(I_j = 1) = \Pr(\text{unit } j \text{ is included in sample given fixed population values for } Y \text{ and the auxiliary variable } X)$. Since we assume that every unit in the population has a non-zero probability of inclusion in the sample, we must have $\pi_j > 0$ for all $j \in U$.

Note that we do NOT assume that the I_j are independent random variables. The properties of the joint distribution of any subset of these random variables will depend on the actual sampling method employed. The simplest joint distribution is of two inclusion variables, I_j and I_k, where $j \neq k$. In this case we put $\pi_{jk} = \Pr_p(I_j = 1,\ I_k = 1) = \Pr(\text{units } j \text{ and } k$ are both included in sample given fixed population values for *Y* and *X*). It is standard to refer to π_j as the inclusion probability for unit *j*, and π_{jk} as the joint inclusion probability for units *j* and *k*.

2.3.1.2 The Horvitz-Thompson estimate

Suppose now that the values π_j are known for each unit in the population. Then, irrespective of which sample is actually chosen, we can define an estimate of *t* of the form

$$\hat{t}_{HT} = \sum_{j\in s} \pi_j^{-1} y_j \; .$$

The notation $j \in s$ means that the summation above is restricted to the sample units, while the subscript *HT* refers to the fact that this estimate was first put forward in Horvitz & Thompson (1952).

2.3.1.3 Design-based theory for the Horvitz-Thompson estimate

It is straightforward to show that the repeated sampling distribution of the sample error of the *HTE* (Horvitz-Thompson estimate) has mean zero. An equivalent way of stating this is to say that $\hat{t}_{HT}$ is unbiased under repeated sampling, or, more commonly, that it is design unbiased – that is, unbiased with respect to repeated sampling under the probability sampling design.

The mean and variance of the repeated sampling distribution of (the sample error defined by) $\hat{t}_{HT}$ are easily obtained. It just requires one to notice that the only random variables contributing to this distribution are the sample inclusion variables I_j defined above. All other quantities (and in particular the values of *Y*) are held fixed at their population values. Consequently, since $\mathrm{E}_p(I_j) = \pi_j$,

$$\begin{aligned} \mathrm{E}_p\left(\hat{t}_{HT} - t\right) &= \mathrm{E}_p\left(\sum_{j\in s} \pi_j^{-1} y_j - \sum_{j\in U} y_j\right) \\ &= \mathrm{E}_p\left(\sum_{j\in U} \frac{I_j y_j}{\pi_j} - \sum_{j\in U} y_j\right) \\ &= \sum_{j\in U} \frac{\mathrm{E}_p(I_j) y_j}{\pi_j} - \sum_{j\in U} y_j = 0 \, . \end{aligned}$$

That is, the sample error distribution of the *HTE* has zero bias under repeated sampling. Note that this proof is dependent on every unit in the population having a non-zero probability of inclusion in sample.

The design variance of $\hat{t}_{HT}$ (that is the variance of the repeated sampling distribution of the sample error defined by $\hat{t}_{HT}$) is obtained through a very similar argument. Since *t* is considered fixed in this case, this variance is given by

$$\begin{aligned}
\mathrm{V}_p\left(\hat{t}_{HT}-t\right) &= \mathrm{V}_p\left(\sum_{J\in s}\pi_j^{-1}y_j\right)=\mathrm{V}_p\left(\sum_{j\in U}\frac{I_j y_j}{\pi_j}\right)\\
&=\sum_{j\in U}\frac{\mathrm{V}_p(I_j)y_j^2}{\pi_j^2}+\sum_{j\in U}\sum_{\substack{k\in U\\k\neq j}}\frac{\mathrm{C}_p(I_j,I_k)y_j y_k}{\pi_j\pi_k}\\
&=\sum_{j\in U}\frac{(1-\pi_j)y_j^2}{\pi_j}+\sum_{j\in U}\sum_{\substack{k\in U\\k\neq j}}\frac{\left(\pi_{jk}-\pi_j\pi_k\right)y_j y_k}{\pi_j\pi_k}.
\end{aligned}$$

Without loss of generality we define $\pi_{jj} = \pi_j$. Then the above variance is

$$\mathrm{V}_p\left(\hat{t}_{HT}-t\right)=\sum_{j\in U}\sum_{k\in U}\frac{\left(\pi_{jk}-\pi_j\pi_k\right)y_j y_k}{\pi_j\pi_k}.$$

Note that this variance is a *FPP*. Consequently we can use the argument that shows $\hat{t}_{HT}$ is design unbiased to obtain an estimate of this variance that is also design unbiased. This is the so-called *HT* estimate of variance

$$\hat{\mathrm{V}}_p^{HT}\left(\hat{t}_{HT}-t\right)=\sum_{j\in s}\sum_{k\in s}\frac{\left(\pi_{jk}-\pi_j\pi_k\right)y_j y_k}{\pi_{jk}\pi_j\pi_k}.$$

2.3.1.4 *Design-based theory for fixed sample size designs*

An important class of sample designs have fixed sample size. For such designs the sum of any realisation of the *N* sample inclusion indicators equals a fixed number *n* (the sample size). It immediately follows that for fixed sample size designs the sum of the population values of π_j must also equal *n*. Furthermore, then

$$\sum_{\substack{k\in U\\k\neq j}}I_j I_k=\left(I_j\sum_{k\in U}I_k\right)-I_j^2=(n-1)I_j\Rightarrow\sum_{\substack{k\in U\\k\neq j}}\pi_{jk}=(n-1)\pi_j.$$

These equalities allow us to express the design variance of $\hat{t}_{HT}$ a little differently. That is, when a fixed sample size design is used this variance is

$$\mathrm{V}_p\left(\hat{t}_{HT}-t\right)=\frac{1}{2}\sum_{j\in U}\sum_{\substack{k\in U\\k\neq j}}\left(\pi_j\pi_k-\pi_{jk}\right)\left(\frac{y_j}{\pi_j}-\frac{y_k}{\pi_k}\right)^2.$$

A design unbiased estimate of this variance is easily seen to be

$$\hat{\mathrm{V}}_p^{SYG}\left(\hat{t}_{HT}-t\right)=\frac{1}{2}\sum_{j\in s}\sum_{\substack{k\in s\\k\neq j}}\left(\frac{\pi_j\pi_k-\pi_{jk}}{\pi_{jk}}\right)\left(\frac{y_j}{\pi_j}-\frac{y_k}{\pi_k}\right)^2.$$

The superscript *SYG* above stands for Sen-Yates-Grundy, the original developers of this particular variance estimate (Yates & Grundy, 1953; Sen, 1953).

The *HT* variance estimate can take negative values when sampled units have high inclusion probabilities. Similarly, the *SYG* variance estimate can be negative if $\pi_j\pi_k < \pi_{jk}$ for some $j \neq k$. Since in most practical cases this condition does not hold, the *SYG* estimate is usually preferred for estimating the design variance of the *HTE*.

2.3.1.5 Approximating second order inclusion probabilities

An important practical problem underlying both variance estimates above is that they require the survey analyst to know the joint inclusion probabilities π_{jk}. In the case of simple random sampling, or stratified random sampling, these probabilities are known. For example, under stratified random sampling

$$\pi_{jk} = \begin{cases} \dfrac{n_h(n_h-1)}{N_h(N_h-1)} & \text{if } j,k \text{ are both in stratum } h; \\ \dfrac{n_h n_g}{N_h N_g} & \text{if } j \text{ is in stratum } h \text{ and } k \text{ is in stratum } g. \end{cases}$$

For other methods of sampling, however, the joint inclusion probabilities are rarely known. In such cases, one can approximate these probabilities so that, within strata, they are at least correct for simple random sampling. That is, we put

$$\pi_{jk} \approx \frac{N_h(n_h-1)}{n_h(N_h-1)}\pi_j\pi_k$$

when j and k are in the same stratum h. Obviously, when j and k are in different strata we have $\pi_{jk} = \pi_j\,\pi_k$.

In the special case of probability proportional to size (*PPS*) sampling Berger (1998) has proposed an alternative approximation. This is based on the following approximation to the variance of the *HTE* (Hajek, 1964):

$$\tilde{\mathrm{V}}_p^H\left(\hat{t}_{HT}-t\right) = \frac{N}{N-1}\left[\sum_{j\in U}\frac{(1-\pi_j)y_j^2}{\pi_j} - d(\pi)G^2(\pi)\right]$$

where

$$d(\pi) = \sum_{j\in U}\pi_j\left(1-\pi_j\right)$$

and

$$G(\pi) = \frac{1}{d(\pi)}\sum_{j\in U}y_j\left(1-\pi_j\right).$$

Berger's variance estimate replaces the population quantities in Hajek's approximation by design-unbiased estimates, leading to the variance estimate

$$\hat{\mathrm{V}}_p^B\left(\hat{t}_{HT}-t\right)=\frac{n\hat{d}(\pi)}{(n-1)d(\pi)}\sum_{j\in s}\left(1-\pi_j\right)\left\{\frac{y_j}{\pi_j}-\hat{G}(\pi)\right\}^2$$

where

$$\hat{d}(\pi)=\sum_{j\in s}\left(1-\pi_j\right)$$

and

$$\hat{G}(\pi)=\frac{1}{\hat{d}(\pi)}\sum_{j\in U}\frac{y_j}{\pi_j}\left(1-\pi_j\right).$$

It should be emphasised that this variance estimator is only suitable for *PPS* designs. It can give seriously misleading results if used with general unequal probability designs. For example, if used with stratified random sampling it has a large positive bias. Conditions for applicability of $\hat{\mathrm{V}}_p^B\left(\hat{t}_{HT}-t\right)$ are set out in Berger (1998).

2.3.1.6 Problems with the design-based approach

The main strength of design-based theory is that it makes no assumptions about the population values being sampled. However this is also its weakness, since there is nothing in the approach to indicate how to make efficient inferences. In particular, the *HTE* can be quite inefficient.

Under the design-based approach to sample survey inference, design unbiasedness is a key measure of quality for a survey estimate. As will be clear from the development above, this property has nothing to do with the actual value of the sample error of this estimate. It is a property of the probability sampling method. On average, over repeated sampling from the fixed finite population of *Y*-values actually “out there”, this error is zero. But the size of the *actual* error may be far from zero. If the variance of the repeated sampling distribution is also small, then this error will be small with high probability. Standard probability theory assures us that this will be the case provided the sample size is “large”. However, there is little to guide one on what “large” means here, since the conditions required for this theory to hold depend on the (unknown) characteristics of the population. Furthermore, in many practical situations sample sizes are NOT “large”, and design-unbiasedness is of limited usefulness. These comments apply equally well to a design-unbiased estimate of the design variance of an estimate. When a sample is not “large” the accuracy of this estimate of variance (that is the difference between it and the true sampling variance of the estimate) suffers from the same problem as the actual sample error itself – we cannot say how small (or how large) it actually is. All we can say is that the procedure used to calculate this estimate will on average produce an estimate that is the right value.

A further problem relates to the use of the design variance as the measure of the error of a particular sample estimate. This quantity is not the actual value of this error. In fact, the design variance remains the same irrespective of the size of this error. This invariance has been criticised (Royall, 1982). Furthermore, the standard estimates of this design variance (which, since they vary from sample to sample, DO vary with the actual error) have been criticised as being misleading. In particular, in some circumstances these variance estimates can be negatively correlated with the actual errors, leading to misleading quality assessments for the survey estimates. See Royall & Cumberland (1981).

Both the above problems (efficient estimates and meaningful variance estimates) can be resolved if one adopts a model-based approach to sample survey inference. However, this is not free of cost. One then has to rely on the adequacy of one's model for the superpopulation distribution of the *Y*-variable of interest. Since all models are, to a greater or lesser extent, incorrect this means that one should adopt robust model-based methods, that is methods that do not seriously lose efficiency under "smooth" deviations from assumptions. This issue is taken up in more detail in 2.3.2.8. Below we develop the basic theory underlining the model-based approach.

2.3.2 The use of models for estimating a population total

As shown above, the design variance of the *HTE* depends on the actual population values of *Y*. Consequently, without some way of "modelling" the distribution of these population *Y*-values, there is little one can say about the properties of the *HTE*. Over the last 25 years a considerable body of theory has therefore developed which attempts to utilise knowledge about the probable distribution of population values for *Y* in order to improve estimation of a *FPP*. Typically, this information is characterised in terms of a stochastic model for this distribution.

There are two basic ways such a model can be used. The *model-assisted* approach essentially uses it to improve estimation of the *FPP* within the design-based framework. That is, the model is used to motivate an estimate with good model-based properties. However, this estimate is still assessed in terms of desirable design-based properties like design unbiasedness and low design variance. Furthermore, the key quality measure of an estimate under this approach remains its estimated design variance.

The other basic approach is fully model-based. Here the restrictions of design unbiasedness and low design variance are dispensed with, being replaced by model unbiasedness and low model variance. Below we describe the basics of the model-based approach. Corresponding development of the model-assisted approach is set out in section 2.3.3.

2.3.2.1 The superpopulation model

In order to describe this approach, we introduce the idea of a superpopulation model. This is a model for the joint distribution of the *N* random variables Y_j, $j \in U$ whose realisations correspond to the population *Y*-values, given the values of the auxiliary variable *X*. Typically such a model specifies the first and second order moments of this joint distribution rather than the complete distribution. Thus we can write

$$\mathrm{E}_\xi(y_j) = \mu(x_j;\omega)$$
$$\mathrm{V}_\xi(y_j) = \sigma^2(x_j;\omega)$$
$$\mathrm{C}_\xi(y_j, y_k) = 0 \quad \text{for } j \neq k$$

where μ and σ are specified functions of x whose values depend on ω, a typically unknown parameter. Note that the assumption that distinct population units are uncorrelated given X may seem restrictive, but is standard for surveys of economic units where X can be quite informative about Y. In household surveys X may provide very little information about Y, in which case it is standard to allow units that "group together" (for example individuals in households) to be correlated. See section 2.3.2.5 below.

2.3.2.2 The homogeneous strata model

This model is widely used in business survey practice. Here, the population is split into strata and it is assumed that the means and variances of the population Y-variables are the same for all units within a stratum, but different across strata. In this case X is a stratum indicator. Assuming the strata are indexed by $h = 1, 2, \ldots, H$, then for j in stratum h we have $\mu(x_j;\omega) = \mu_h$ and $\sigma(x_j;\omega) = \sigma_h$. *Note that this model does not assume any relationship between the stratum means and variances*.

2.3.2.3 The simple linear regression model

Another commonly used model is where x_j is a measure of the "size" of the j^{th} population unit, and it is reasonable to assume a linear relationship between Y and X. Typically this linear relationship is coupled with heteroskedasticity in X, in the sense that the variability in Y tends to increase with increasing X. A specification that allows for this behaviour for positive valued X is $\mu(x_j;\omega) = \alpha + \beta x_j$ and $\sigma(x_j;\omega) = \psi + \phi x_j^\gamma$. In many economic populations the regression of Y on X goes through the origin, and this model reduces to the simple "ratio" form defined by $\alpha = \psi = 0$.

2.3.2.4 The general linear regression model

Both the homogeneous strata model and the simple linear regression model are special cases of a model where the auxiliary information corresponding to X contains a mix of stratum identifiers and size variables. We denote this multivariate auxiliary variable by **X**. Then $\mu(\mathbf{x}_j;\omega) = \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}$. It is standard in this case to express the heteroskedasticity in Y in terms of a single auxiliary variable Z, which can be one of the auxiliary size variables in **X**, or some positive valued function of the components of this vector (for example a power transformation like x^γ above). In either case we put $\sigma(\mathbf{x}_j;\omega) = \sigma z_j$. It is important to note that the specification of **X** is quite general. In most applications this vector contains only "main effects", but conceptually there is nothing to stop it containing any function (including interaction terms) defined by the auxiliary information on the sample frame.

2.3.2.5 The cluster model

A common feature of the models set out above is that they assume individual population units are uncorrelated, irrespective of their "distance" from other population units. That is, after conditioning on the auxiliary information in X, there is no reason to expect population units that are contiguous in some sense to be "more alike" with respect to their values of Y than units that are not contiguous. Another way of expressing this is that these models assume the observed similarity in Y values for contiguous units is completely explained by their similar values of X.

When the explanatory power of X is weak, as is the case in most human populations, this assumption of lack of correlation cannot be sustained. In such cases it is usual to expand the model in 2.3.2.1 to allow correlation between contiguous units. In particular, a hierarchical structure for the population is often assumed, with individuals grouped together into small non-overlapping clusters (for example households). All clusters are assumed to be more or less similar in size, and essentially similar in terms of the range of Y values they contain. However, individuals from the same cluster are assumed to be more alike than individuals from different clusters. Typically this is modelled by an unobservable "cluster effect" variable which has a distribution across the clusters making up the population. The effect of this variable is to induce a within cluster correlation for Y.

Since the focus of this report is quality measures for business surveys, and cluster type models are rarely used to model business populations, we will not pursue this issue any further. See Royall (1986) for further discussion of model-based estimation under a cluster specification.

2.3.2.6 Ignorable sampling

An important assumption that is typically made at this stage is that the joint distribution of the *sample* values of Y can be deduced from the assumed superpopulation model. In particular, it is often assumed that if unit j is in sample, then the mean and variance of y_j are the same as specified by the model. That is, the fact that a unit is selected in the sample has no impact on our uncertainty about the distribution of potential values associated with its corresponding Y-value. This is the so-called *ignorable sampling* assumption. It is satisfied by any method of probability sampling that depends at most on known population auxiliary information. We shall assume ignorable sampling in what follows, since this is what is done in practice. An investigation of non-ignorable sampling is set out in Chapter 4.

2.3.2.7 Bias, variance and mean squared error under the model-based approach

Under the model-based approach the total t of the population values of Y is a random variable, so the problem of estimating this *FPP* is actually a *prediction* problem. An estimate $\hat{t}$ of the population total of Y is a function of the sample Y-values, each one of which is a realisation of a random variable under the assumed superpopulation model. Consequently $\hat{t}$ is also the realisation of a random variable. The sample error $\hat{t}-t$ is a prediction error under this approach. The *model bias* of an estimate $\hat{t}$ of t is then the expected value of its sample error under the model, that is $\mathrm{E}_{\xi}(\hat{t}-t)$. This estimate is said to be *model unbiased* if this model bias is zero, that is $\mathrm{E}_{\xi}(\hat{t}-t)=0$.

The *model mean squared error* of $\hat{t}$ is the sum of its *model variance* and the square of its model bias

$$\mathrm{E}_{\xi}(\hat{t}-t)^2=\mathrm{V}_{\xi}(\hat{t}-t)+\left[\mathrm{E}_{\xi}(\hat{t}-t)\right]^2.$$

Note that both bias and mean squared error above will depend on ω. Provided this parameter can be estimated from the sample data, say by $\hat{\omega}$, then we can estimate the model mean squared error of $\hat{t}$ by replacing ω by $\hat{\omega}$ in the variance and bias terms above. Such a "plug-in" estimate may itself be biased, however. Bias corrections can be constructed, depending on the actual population model assumed.

2.3.2.8 Weaknesses of the model-based approach

It is important to realise that the model-based properties of an estimate are a consequence of the superpopulation model assumed. Since the "correctness" of this assumption is essentially unverifiable (although the sample data can throw light on its appropriateness) there has been criticism of this approach as being *model dependent*. A crucial quality requirement of a model-based approach therefore is *robustness* to specification of the superpopulation model.

There are two basic ways such robustness can be achieved. The first (and most effective) is to design the sample so that the survey estimate is in fact model unbiased with respect to both the superpopulation model thought to be most appropriate for the population values as well as with respect to a large class of alternative superpopulation models that could potentially underlie these values. The second (and typically less effective) is to use a very general model, typically one that is overspecified, at the estimation stage of the survey. That is, we replace the original survey estimate (which was designed to be unbiased with respect to a much "smaller" model) by the estimate suggested by this extended model. See section 2.3.4.

Probability sampling is a key element of a robust sample design strategy. This is because probability sampling can provide *average robustness* by selecting samples where the bias due to misspecification of the superpopulation model is small. However, it is usually advised that one should not rely entirely on probability sampling in this regard, effectively leaving robustness "to chance", but that one should also implement robust sample design strategies like size stratification and ordered systematic sampling within strata. These strategies effectively "spread" the sample across the population in such a way that misspecification bias is considerably reduced. For further discussion of this issue see Royall & Herson (1973).

2.3.2.9 Linear prediction

A widely used class of estimates of *t* is linear in the sample *Y*-values. That is, the estimate $\hat{t}$ is of the form

$$\hat{t}_L=\sum_{j\in s}w_{js}y_j\ .$$

In general, the weight w_{js} above will depend on x_j and will be sample dependent, in the sense that it will also depend on the *X*-values of *all* the sample units. However, it is not a function of the sample *Y*-values, and hence is a fixed quantity under the population model. This is in contrast to the design-based approach, which would treat this weight as a random variable in this case.

For a large number of commonly used superpopulation models it is possible to construct weights w_{js} that ensure the linear estimate $\hat{t}_L$ above is model unbiased and has minimum prediction variance. Such weights are typically referred to as Best Linear Unbiased (*BLU*) sample weights, and the estimate $\hat{t}_L$ is then the Best Linear Unbiased Predictor (*BLUP*) of *t* under the model. Since these weights depend on the actual superpopulation model, they will vary according to how this model is specified.

To illustrate, the homogeneous strata model and the linear regression model of 2.3.2.2 and 2.3.2.3 are often merged to give a model where the population is partitioned into strata with a separate regression relationship between the study variable *Y* and the auxiliary variable *X* in each stratum. If, in addition, both the linear regression of *Y* on *X* in each stratum, and the variation of *Y* about this regression line, are strictly proportional to *X* (that is the regression line goes through the origin, with residual variance proportional to *X*) then the general model in 2.3.2.1 becomes

$$\begin{aligned} \mathrm{E}_\xi(y_j) &= \beta_h x_j \quad && \text{for } j \in \text{stratum } h \\ \mathrm{V}_\xi(y_j) &= \sigma_h^2 x_j \quad && \text{for } j \in \text{stratum } h \\ \mathrm{C}_\xi(y_j, y_k) &= 0 \quad && \text{for } j \neq k \end{aligned}$$

Under this model the *BLUP* of *t* is the *separate ratio estimator*

$$\hat{t}_{R,sep} = \sum_h N_h \hat{b}_h \bar{x}_h = \sum_h N_h \bar{y}_{sh} (\bar{x}_h / \bar{x}_{sh})$$

where $\hat{b}_h$ is the Best Linear Unbiased Estimate (*BLUE*) of β_h, defined as the ratio of the sample mean $\bar{y}_{sh}$ of *Y* in stratum *h* to the corresponding sample mean $\bar{x}_{sh}$ of *X*, and $\bar{x}_h$ is the population mean of *X* in stratum *h*. Note that this estimator is a particular case of $\hat{t}_L$, with $w_{js} = (N_h \bar{x}_h)/(n_h \bar{x}_{sh})$ for sample unit *j* in stratum *h*.

Under the superpopulation model set out in 2.3.2.1, the model bias of $\hat{t}_L$ is easily seen to be

$$\mathrm{E}_\xi(\hat{t}_L - t) = \sum_{j \in s} w_{js} \mu(x_j; \omega) - \sum_{j \in U} \mu(x_j; \omega)$$

Since $\mu(x_j; \omega)$ is *O*(1), it immediately follows that w_{js} must be *O*(*N*/*n*) if $\hat{t}_L$ is to be model unbiased. Furthermore, the prediction variance of $\hat{t}_L$ under the model is

$$\mathrm{V}_\xi(\hat{t}_L - t) = \sum_{j \in s} (w_{js} - 1)^2 \sigma^2(x_j; \omega) + \sum_{j \notin s} \sigma^2(x_j; \omega).$$

Given an estimate $\hat{\omega}$ of ω calculated from the sample data, a simple "plug-in" estimate of this prediction variance is

$$\hat{\mathrm{V}}_{\xi}^{L}\left(\hat{t}_L - t\right) = \sum_{j\in s}\left(w_{js} - 1\right)^2 \sigma^2(x_j;\hat{\omega}) + \sum_{j\notin s}\sigma^2(x_j;\hat{\omega})$$

Since w_{js} is $O(N/n)$, it is clear that the leading term in this estimated variance is the first (sample) term on the right hand side above. The validity of this estimate therefore rests on the accuracy of $\sigma^2\left(x_j;\omega\right)$ as a specification for the variance of y_j.

Returning to the case of the separate ratio estimator defined above, one can show that this estimated prediction variance then becomes

$$\hat{\mathrm{V}}_{\xi}^{L}\left(\hat{t}_{R,sep} - t\right) = \sum_{h}\hat{\sigma}_h^2\left\{\frac{\left(N_h\bar{x}_h\right)^2}{n_h\bar{x}_{sh}}\left(1 - \frac{n_h\bar{x}_{sh}}{N_h\bar{x}_h}\right)\right\}$$

where

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1}\sum_{j\in sh}\left(y_j - \hat{b}_h x_j\right)^2 / x_j \, .$$

2.3.2.10 Robust prediction variance estimation

A more robust estimate of the prediction variance of $\hat{t}_L$ can be defined by replacing this leading term by one whose validity only depends on the superpopulation model being correct to first, rather than second, order. In particular, suppose $\hat{\mu}_j = \mu\left(x_j;\hat{\omega}\right)$ is an unbiased estimate of $\mu\left(x_j;\omega\right)$ under the superpopulation model. Then

$$\mathrm{E}_{\xi}\left(y_j - \hat{\mu}_j\right)^2 = \mathrm{V}_{\xi}(y_j) + O\left(n^{-1}\right)$$

irrespective of the actual "true" specification of the superpopulation variance of y_j. Consequently the alternative prediction variance estimate for $\hat{t}_L$

$$\hat{\mathrm{V}}_{\xi}^{R}\left(\hat{t}_L - t\right) = \sum_{j\in s}\left(w_{js} - 1\right)^2\left(y_j - \hat{\mu}_j\right)^2 + \sum_{j\notin s}\sigma^2\left(x_j;\hat{\omega}\right)$$

will be valid even when the second order moments in the superpopulation model are incorrectly specified. In practice, slightly modified versions of this robust variance estimate are usually employed, typically with the squared residual above multiplied by an $O(1)$ adjustment, thus ensuring it is also then an unbiased estimate of the variance of y_j as specified by the superpopulation model.

To illustrate this approach, consider the case where it is convenient to assume that all units in some specified part of the population (for example a stratum) have the same mean value, say μ, and the same variance σ^2. For convenience we shall assume that this subpopulation is the only one we are interested in, and so we treat it as the target population. Suppose also that sample weights w_{js} are available, and it is proposed to estimate t using the linear predictor $\hat{t}_L$ described in 2.3.2.9. Under this model

$$\mathrm{E}_\xi\left(\hat{t}_L - t\right) = \mu\left(\sum_{j\in s} w_{js} - N\right)$$

so the sample weights have to sum to the population size N for this estimate to be unbiased. We assume this. The prediction variance of $\hat{t}_L$ is then

$$\mathrm{V}_\xi\left(\hat{t}_L - t\right) = \sigma^2\left(\sum_{j\in s}\left(w_{js} - 1\right)^2 + (N-n)\right).$$

An unbiased estimate of μ is the weighted average

$$\hat{\mu}_w = N^{-1}\sum_{j\in s} w_{js} y_j \ .$$

Also, an unbiased estimate of σ^2 is then

$$\hat{\sigma}_w^2 = \frac{1}{n}\sum_{j\in s}\left(1 - 2\frac{w_{js}}{N} + \frac{1}{N^2}\sum_{k\in s} w_{ks}^2\right)^{-1}\left(y_j - \hat{\mu}_w\right)^2$$

so an unbiased estimate of the prediction variance of $\hat{t}_L$ under this model is

$$\hat{\mathrm{V}}_\xi\left(\hat{t}_L - t\right) = \hat{\sigma}_w^2\left(\sum_{j\in s}\left(w_{js} - 1\right)^2 + (N-n)\right).$$

Unfortunately, this estimate will be biased if the assumption of constant variance for the y_j is incorrect. In particular, suppose that the units in the population have potentially different (and unknown) variances, say σ_j^2. To distinguish this case from the constant variance model ξ assumed so far, we use a subscript of η below. The *true* prediction variance of $\hat{t}_L$ will then be

$$\mathrm{V}_\eta\left(\hat{t}_L - t\right) = \sum_{j\in s}\left(w_{js} - 1\right)^2\sigma_j^2 + \sum_{j\notin s}\sigma_j^2 \ .$$

The robust variance estimate $\hat{\mathrm{V}}_\xi^R$ is then

$$\hat{\mathrm{V}}_\xi^R\left(\hat{t}_L - t\right) = \sum_{j\in s}\left(w_{js} - 1\right)^2\left(y_j - \hat{\mu}_w\right)^2 + (N-n)\hat{\sigma}_w^2$$

It is easy to see that this robust variance estimate will not be exactly unbiased under ξ. However, the slightly modified alternative $\hat{\mathrm{V}}_\xi^D$ below is:

$$\hat{\mathrm{V}}_\xi^D\left(\hat{t}_L - t\right) = \sum_{j\in s}\left(\frac{\left(w_{js} - 1\right)^2\left(y_j - \hat{\mu}_w\right)^2}{1 - 2\frac{w_{js}}{N} + \frac{1}{N^2}\sum_{k\in s} w_{ks}^2}\right) + (N-n)\hat{\sigma}_w^2 \ .$$

Extension of this robust approach to prediction variance estimation for the separate ratio estimator introduced in 2.3.2.9 is discussed in Royall & Cumberland (1981). This leads to the variance estimate

$$\hat{\mathrm{V}}_{\xi}^{D}(\hat{t}_{R,sep}-t)=\sum_{h}\left(\frac{N_h^2}{n_h}\right)\left(\frac{\bar{x}_h}{\bar{x}_{sh}}-\frac{n_h}{N_h}\right)\left(\frac{\bar{x}_h}{\bar{x}_{sh}}\right)\frac{1}{n_h}\sum_{j\in sh}\left\{\frac{(y_j-\hat{b}_h x_j)^2}{1-\left(x_j/n_h\bar{x}_{sh}\right)}\right\}$$

$$=\sum_{h}\left(\frac{N_h^2}{n_h}\right)\left(\frac{\bar{x}_h}{\bar{x}_{sh}}\right)^2\frac{1}{n_h}\sum_{j\in sh}\left\{\frac{(y_j-\hat{b}_h x_j)^2}{1-\left(x_j/n_h\bar{x}_{sh}\right)}\right\}-\text{lower order term.}$$

As we shall see in 2.3.3.2 below, it turns out that the leading term above in this robust model-based variance estimate is essentially identical to a design-based variance estimate for the separate regression estimate that arises under the model-assisted approach to sample survey inference.

2.3.3 The model-assisted approach

An alternative approach to incorporating superpopulation model information into survey estimation is to use the model to suggest improvements to the standard *HTE*, but to continue to base all inference on the design-based properties of the resulting estimate. This approach is commonly referred to as model-assisted. See Särndal *et al.* (1992).

2.3.3.1 The GREG and GRAT estimates for a population total

Given a superpopulation model of the form set out in 2.3.2.1, there are two standard ways the *HTE* is typically "improved upon". This is via generalised regression estimation (*GREG*) or via generalised ratio estimation (*GRAT*). In order to motivate these approaches, consider the following equivalent ways of rewriting the population total *t* of *Y*, where $\mu_j=\mu(x_j;\omega)$,

$$t=\sum_{j\in U}\mu_j+\sum_{j\in U}(y_j-\mu_j)=\sum_{j\in U}\mu_j+\sum_{j\in U}e_j$$

and

$$t=\sum_{j\in U}\mu_j\frac{\sum_{j\in U}y_j}{\sum_{j\in U}\mu_j}=R\sum_{j\in U}\mu_j\,.$$

An improved estimate of *t* based on the first decomposition above can then be defined by replacing the unknown μ_j by a suitably chosen "plug-in" estimate, and the population total of the e_j by its *HTE*. This leads to the *GREG* extension of the *HTE*:

$$\hat{t}_{GREG}=\sum_{j\in U}\hat{\mu}_j+\sum_{j\in s}\frac{\hat{e}_j}{\pi_j}$$

where $\hat{\mu}_j=\mu(x_j;\hat{\omega}_p)$, $\hat{e}_j=y_j-\hat{\mu}_j$ and $\hat{\omega}_p$ is a "design consistent" estimate of the parameter ω defined by the superpopulation model. Typically, the last condition is equivalent to requiring that in large populations and samples, $\hat{\omega}_p$ has a design bias of $O\left(n^{-1/2}\right)$ when used as an estimate of a *FPP* ω_N, which is itself a model unbiased estimate of ω based on the full population.

An alternative improved estimate of *t* can be based on the second decomposition above. This is the *GRAT* extension of the *HTE*:

$$\hat{t}_{GRAT} = \left(\sum_{j\in s}\frac{y_j}{\pi_j}\right)\left(\sum_{j\in s}\frac{\hat{\mu}_j}{\pi_j}\right)^{-1}\sum_{j\in U}\hat{\mu}_j = \hat{R}_p\sum_{j\in U}\hat{\mu}_j \ .$$

Clearly the design unbiasedness of the *HTE*, coupled with the design consistency of $\hat{\omega}_p$, ensures that both the *GREG* and the *GRAT* are approximately design unbiased in large samples.

2.3.3.2 Variance estimates for the GREG and GRAT

Exact expressions for the design variances of the *GREG* and *GRAT* estimates are unavailable in general. However, it is relatively straightforward to write down first order approximations. In the case of the *GREG*, one can note that the design consistency of $\hat{\omega}_p$ implies that the leading term in the design variance of this estimate is the design variance of the generalised difference "estimate" $\tilde{t}_{GDIFF}$, which is just the *GREG* estimate but with $\hat{\omega}_p$ replaced by ω_N. The *HT* estimate of variance for this generalised difference estimate is

$$\hat{\mathrm{V}}_p^{HT}\left(\hat{t}_{GDIFF} - t\right) = \sum_{j\in s}\sum_{k\in s}\frac{(\pi_{jk} - \pi_j\pi_k)}{\pi_{jk}\pi_j\pi_k}\left(y_j - \mu(x_j;\omega_N)\right)\left(y_k - \mu(x_k;\omega_N)\right).$$

On the other hand, if a fixed sample size design has been used, the *SYG* variance estimate can be calculated

$$\hat{\mathrm{V}}_p^{SYG}(\tilde{t}_{GDIFF} - t) = \frac{1}{2}\sum_{j\in s}\sum_{\substack{k\in s\\ k\neq j}}\left(\frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(\frac{y_j - \mu(x_j;\omega_N)}{\pi_j} - \frac{y_k - \mu(x_k;\omega_N)}{\pi_k}\right)^2 .$$

A first order estimate of the design variance of the *GREG* is then obtained by substituting $\hat{\omega}_p$ for ω_N in either of the above variance estimates. For example the *SYG* estimate of the design variance of the *GREG* is

$$\hat{\mathrm{V}}_p(\hat{t}_{GREG} - t) = \frac{1}{2}\sum_{j\in s}\sum_{\substack{k\in s\\ k\neq j}}\left(\frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(\frac{\hat{e}_j}{\pi_j} - \frac{\hat{e}_k}{\pi_k}\right)^2 .$$

A similar leading term approximation to the design variance of the *GRAT* can be developed. We again replace $\hat{\omega}_p$ by ω_N in the specification of this estimate and then use a first order

Taylor Series approximation to the variance of the ratio term in the resulting "estimate" to get the approximation (C_p denotes design-based covariance)

$$\mathrm{V}_p(\tilde{t}_{GRAT} - t) \approx \mathrm{V}_p\left(\sum_{j\in s}\frac{y_j}{\pi_j}\right) - 2R_N\,\mathrm{C}_p\left(\sum_{j\in s}\frac{y_j}{\pi_j}, \sum_{j\in s}\frac{\mu(x_j;\omega_N)}{\pi_j}\right) + R_N^2\,\mathrm{V}_p\left(\sum_{j\in s}\frac{\mu(x_j;\omega_N)}{\pi_j}\right)$$

where

$$R_N = \frac{\sum_{j\in U} y_j}{\sum_{j\in U}\mu(x_j;\omega_N)}.$$

Assuming a fixed sample size design and substituting *SYG* estimates for the variances and covariances on the right hand side of this expression, replacing ω_N by $\hat{\omega}_p$, R_N by $\hat{R}_p$ and collecting terms leads to the following first order estimate for the design variance of the *GRAT*

$$\hat{\mathrm{V}}_p(\tilde{t}_{GRAT} - t) = \frac{1}{2}\sum_{j\in s}\sum_{\substack{k\in s\\ k\neq j}}\left(\frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(\frac{y_j - \hat{R}_p\hat{\mu}_j}{\pi_j} - \frac{y_k - \hat{R}_p\hat{\mu}_k}{\pi_k}\right)^2 .$$

Note that this estimate is similar to, but not the same as, the variance estimate for the *GREG*.

If the mean function $\mu(x;\ \omega)$ is linear in x, and the estimate $\hat{\omega}_p$ is a model-unbiased linear function of the sample Y-values, then both the *GREG* and *GRAT* estimators are also model-unbiased linear functions of the sample Y-values. That is, they can be written in the form $\hat{t}_L$ introduced in 2.3.2.9. In such a case we can derive an alternative variance estimate for the *GREG*/*GRAT* which is closely related to the robust model-based prediction variance estimates described in 2.3.2.10.

To start, put $\tilde{\mu}_j = \mu\left(x_j;\omega_N\right)$ and $\tilde{e}_j = y_j - \tilde{\mu}_j$. Let w_{js} denote the sample weight of the j^{th} sample unit in the "linear representation" of the *GREG*. Since the mean function is linear in x, and the *GREG* is model-unbiased, it immediately follows that

$$\sum_{j\in U}\tilde{\mu}_j = \sum_{j\in s} w_{js}\tilde{\mu}_j .$$

Consequently the *GREG* can be equivalently written

$$\hat{t}_{GREG} = \sum_{j\in U}\tilde{\mu}_j + \sum_{j\in s} w_{js}\tilde{e}_j = \sum_{j\in U}\tilde{\mu}_j + \sum_{j\in s}\frac{g_{js}\tilde{e}_j}{\pi_j}$$

where $g_{js} = w_{js}\pi_j$ is the g-weight associated with the *GREG*. It immediately follows that

$$\mathrm{V}_p\left(\hat{t}_{GREG} - t\right) = \mathrm{V}_p\left(\sum_{j\in s}\frac{g_{js}\tilde{e}_j}{\pi_j}\right) = \sum_{j\in U}\sum_{k\in U}\frac{\left(\pi_{jk} - \pi_j\pi_k\right)g_{js}\tilde{e}_j g_{ks}\tilde{e}_k}{\pi_j\pi_k}$$

and we can use standard design-based theory to write down an estimate of this variance, substituting $\hat{e}_j = y_j - \hat{\mu}_j$ for the unknown $\tilde{e}_j$. For example, the *HT* estimate of variance arising from this representation is

$$\hat{\mathrm{V}}_p^{HT}\left(\hat{t}_{GREG} - t\right) = \sum_{j \in s}\sum_{k \in s}\frac{\left(\pi_{jk} - \pi_j \pi_k\right)}{\pi_{jk}\pi_j\pi_k} g_{js}\hat{e}_j g_{ks}\hat{e}_k$$

while the *SYG* version is

$$\begin{aligned}\hat{\mathrm{V}}_p^{SYG}\left(\tilde{t}_{GREG} - t\right) &= \frac{1}{2}\sum_{j \in s}\sum_{\substack{k \in s \\ k \neq j}}\left(\frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(\frac{g_{js}\hat{e}_j}{\pi_j} - \frac{g_{ks}\hat{e}_k}{\pi_k}\right)^2 \\ &= \frac{1}{2}\sum_{j \in s}\sum_{\substack{k \in s \\ k \neq j}}\left(\frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(w_{js}\hat{e}_j - w_{ks}\hat{e}_k\right)^2.\end{aligned}$$

Equivalent variance estimates for the *GRAT* are easily developed.

We illustrate the preceding theory by returning to the case of the separate ratio estimate introduced in 2.3.2.9, assuming in addition that the sampling method within a stratum is simple random sampling. Since sample inclusion probabilities within a stratum are constant under this design, and recollecting that the definition of $\hat{b}_h$ ensures the sum of residuals within a stratum is zero, we can represent this estimate in the form

$$\hat{t}_{R,sep} = \sum_h \sum_{j \in U_h}\hat{b}_h x_j = \sum_h \sum_{j \in U_h}\hat{b}_h x_j + \sum_h \sum_{j \in s_h}\frac{y_j - \hat{b}_h x_j}{\pi_j}.$$

That is, the separate ratio estimate is a *GREG* estimate, with *g*-weight $g_{js} = \bar{x}/\bar{x}_{sh}$ for sample unit *j* in stratum *h*. Furthermore, under simple random sampling within a stratum the *HT* and *SYG* variance estimates are identical, and so the theory above leads to the variance estimate

$$\begin{aligned}\hat{\mathrm{V}}_p^{HT}\left(\hat{t}_{R,sep} - t\right) &= \sum_h \left(\frac{\bar{x}_h}{\bar{x}_{sh}}\right)^2 \sum_{j \in s_h}\sum_{k \in s_h}\frac{\left(\pi_{jk} - \pi_j\pi_k\right)}{\pi_{jk}\pi_j\pi_k}\left(y_j - \hat{b}_h x_j\right)\left(y_k - \hat{b}_h x_k\right) \\ &= \sum_h \left(\frac{N_h^2}{n_h}\right)\left(\frac{\bar{x}_h}{\bar{x}_{sh}}\right)^2 \frac{1}{n_h - 1}\sum_{j \in s_h}\left(y_j - \hat{b}_h x_j\right)^2.\end{aligned}$$

As noted at the end of 2.3.2.10, this is essentially the leading term in the robust model-based variance estimate for the separate ratio estimate.

2.3.4 Calibration weighting

This is an area of survey estimation that has seen considerable development over the last five years. It is also an area where both design-based and model-based ideas are relevant. Basically, calibration is the process by which a set of survey weights (either model-based *BLU* weights or design-based inverse π-weights) are modified in a "minimal" way so that when these modified weights are applied to specified “control” variables, known population totals for these variables are recovered from the survey data.

Design-based justification for calibration is mainly heuristic. The idea is that since the calibrated weights recover population control totals, they should also be “good” for other survey variables. Calibration makes more sense from a model-assisted viewpoint, since with certain types of calibration (essentially based on a minimum chi-square criterion for the “distance” between the original uncalibrated weights and the calibrated weights), calibration is equivalent to *GREG* estimation based on a superpopulation model that is linear in the variables defining the control totals. From a model-based viewpoint minimum chi-square calibration is straightforward. It essentially corresponds to modifying the initial set of sample weights so that the final calibrated estimate is model unbiased under this linear superpopulation model. Other types of distance criteria can be similarly model-motivated.

In the model-based framework calibration is a natural way to generalise sample weights so they are valid under “larger” models (specified by the control totals) than those that were originally thought to be appropriate for the population. In this sense calibration is also a strategy for dealing with a *multipurpose* survey, particularly one with many *Y* variables each one following perhaps a different superpopulation model specified by different *X*-variables. By calibrating to the control totals of each of these potential covariates, one can define a single sample weight that should lead to unbiased estimates for any particular *Y* variable.

Since choice of calibration control totals is equivalent to choice of a superpopulation model, all the problems associated with under- and over-specification of such models flow through to calibration weighting. Thus calibrating on too large a range of control totals is analogous to model *overspecification* and tends to result in inefficient estimates and highly variable weights. In particular, under minimum chi-square calibration one can obtain weights that are negative or large positive in such cases. On the other hand, missing out a key calibration constraint is equivalent to leaving a key explanatory factor out of a model, and can lead to substantial model bias in the survey estimate.

Since assessing the quality of survey weighting methodology is not the primary focus of this report, we do not pursue this issue further. Interested readers are referred to Chambers (1997).

2.4 Methods for nonlinear functions of the population values

Although estimation of population totals is a key objective of many business surveys, it is also important to be able to construct estimates of *FPP*’s that are nonlinear functions of the population values. For example, ratios of population totals are often of interest, as are finite population quantiles.

2.4.1 Variance estimation via Taylor series linearisation

2.4.1.1 Differentiable functions of population totals

In general, let $\theta = f(t_1, t_2, \ldots, t_m)$ denote a differentiable function of the population totals of *m* *Y*-variables. Furthermore, let $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m$ denote estimates of these totals. A natural estimate of θ is then the "plug-in" estimate $\hat{\theta} = f(\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m)$. If the component estimates $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m$ are unbiased, then $\hat{\theta}$ will be approximately unbiased in large samples.

A first order approximation to the sample error of $\hat{\theta}$ is

$$\hat{\theta} - \theta = f(\hat{t}_1, \ldots, \hat{t}_m) - f(t_1, \ldots t_m) \approx \sum_{a=1}^{m} \frac{\partial f}{\partial t_a}(\hat{t}_a - t_a)$$

where $\partial f / \partial t_a$ denotes the partial derivative of *f* with respect to its a^{th} argument, evaluated at $t_1, t_2, \ldots, t_m$. Consequently, under either the design-based or model-based approaches, a first order approximation to the variance of this sample error is

$$\mathrm{V}(\hat{\theta} - \theta) \approx \sum_{a=1}^{m} \sum_{b=1}^{m} \left(\frac{\partial f}{\partial t_a} \right) \left(\frac{\partial f}{\partial t_b} \right) \mathrm{C}(\hat{t}_a - t_a, \hat{t}_b - t_b).$$

Here V denotes variance and C denotes covariance. It immediately follows that an estimate of this first order approximation is

$$\hat{\mathrm{V}}(\hat{\theta} - \theta) \approx \sum_{a=1}^{m} \sum_{b=1}^{m} \left(\frac{\partial f}{\partial \hat{t}_a} \right) \left(\frac{\partial f}{\partial \hat{t}_b} \right) \hat{\mathrm{C}}(\hat{t}_a - t_a, \hat{t}_b - t_b)$$

where $\hat{\mathrm{C}}$ denotes an estimated covariance and $\partial f / \partial \hat{t}_a$ denotes the partial derivative of *f* with respect to its a^{th} argument, evaluated at $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m$. Note that $\hat{\mathrm{C}}$ can be calculated using any of the different variance estimation methods described in section 2.3.

An important special case is where the estimates $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m$ all have the linear form discussed in 2.3.2.9, which includes the *HTE*, linear prediction estimation and calibration estimation. Then straightforward algebra can be used to show

$$\mathrm{V}(\hat{\theta} - \theta) \approx \mathrm{V}(\hat{t}_z - t_z)$$

where t_Z is the population total of the linearised variable

$$z_j = \sum_{a=1}^{m} \left(\frac{\partial f}{\partial t_a} \right) y_{aj}$$

and $\hat{t}_z$ is the linear weighted estimate of this total. That is,

$$\hat{t}_z = \sum_{j \in s} w_{js} z_j = \sum_{j \in s} w_{js} \left(\sum_{a=1}^{m} \left(\frac{\partial f}{\partial t_a} \right) y_{aj} \right).$$

Note that y_{aj} denotes the value of the variable defining t_a for the j^{th} population unit. In principle a first order approximation to the variance of $\hat{\theta}$ can then be computed as the estimated variance of the sample error of $\hat{t}_z$.

In practice we do not know the values of the partial derivatives defining z_j since they are evaluated at the unknown t_a, $a = 1,2,\ldots,m$. However these values can be replaced by the estimates $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m$, to give an estimate $\hat{z}_j$ which replaces z_j in the formula for $\hat{t}_z$ above and is then treated as a "standard" *Y*-variable. This approach was first suggested by Woodruff (1971).

2.4.1.2 Functions defined as solutions of estimating equations

Not all *FPP*'s of interest can be expressed as smooth functions of the population totals of distinct *Y*-variables, for example the finite population median. A wider class of *FPP*'s is therefore obtained by considering those that can be defined as solutions to population level estimating equations. In general, θ is defined by a population level estimating equation if it is a solution to

$$H(\theta) = \sum_{j \in U} f(y_{1j}, \ldots, y_{mj}; \theta) = 0$$

where *f* is typically assumed to be a differentiable function of θ. A "linear" estimate of θ is $\hat{\theta}$, where

$$\hat{H}(\hat{\theta}) = \sum_{j \in s} w_{js} f(y_{1j}, \ldots, y_{mj}; \hat{\theta}) = 0 .$$

Taylor series linearisation can be used estimate the variance of $\hat{\theta}$. We write

$$0 = \hat{H}(\hat{\theta}) \approx \hat{H}(\theta) + \left(\frac{\partial \hat{H}}{\partial \theta}\right)(\hat{\theta} - \theta) = \hat{H}(\theta) + (\hat{\theta} - \theta) \sum_{j \in s} w_{js} \frac{\partial f(y_{1j}, \ldots, y_{mj}; \theta)}{\partial \theta}$$

from which we obtain the first order approximation

$$\mathrm{V}(\hat{\theta} - \theta) = \left(\sum_{j \in s} w_{sJ} \frac{\partial f(y_{1j}, \ldots, y_{mj}; \theta)}{\partial \theta}\right)^{-1} \mathrm{V}(\hat{H}(\theta)) \left(\sum_{j \in s} w_{js} \frac{\partial f(y_{1j}, \ldots, y_{mj}; \theta)}{\partial \theta}\right)^{-1} .$$

The so-called "sandwich" estimate of variance is obtained by evaluating the partial derivatives above at $\hat{\theta}$, and replacing the variance term in the middle by an appropriate "plug-in" estimate. For arbitrary θ

$$\mathrm{V}(\hat{H}(\theta)) = \mathrm{V}\left(\sum_{j \in s} w_{js} f(y_{1j}, \cdots y_{mj}; \theta)\right) = \mathrm{V}\left(\sum_{j \in s} w_{js} z_j(\theta)\right)$$

where $z_j(\theta) = f(y_{1j}, \cdots y_{mj}; \theta)$ is just another population *Y*-variable. Consequently the variance on the right hand side above is the variance of a linear estimate of the population total of this derived variable, and we can use the theory developed in the previous section to estimate it. “Plugging in” $\hat{\theta}$ for θ in this variance estimate gives an estimate of this variance when θ is replaced by $\hat{\theta}$. We denote this estimate by $\hat{\mathrm{V}}\left(\hat{H}\left(\hat{\theta}\right)\right)$. The final sandwich estimate of variance for $\hat{\theta}$ is then

$$\hat{\mathrm{V}}\left(\hat{\theta}-\theta\right)=\left(\sum_{j\in s} w_{js}\frac{\partial f\left(y_{1j},\cdots y_{mj};\hat{\theta}\right)}{\partial\hat{\theta}}\right)^{-1}\left(\hat{\mathrm{V}}\left(\hat{H}\left(\hat{\theta}\right)\right)\right)\left(\sum_{j\in s} w_{js}\frac{\partial f\left(y_{1j},\cdots y_{mj};\hat{\theta}\right)}{\partial\hat{\theta}}\right)^{-1}.$$

2.4.2 Replication-based methods for variance estimation

Although most *FPP*’s of interest can be defined in terms of a smooth function of population totals, or as the solution of a population estimating equation, there remain situations where the definition of the *FPP* is so complex that application of Taylor series linearisation methods for variance estimation is difficult. In such cases we can use alternative variance estimation methods that are “simple” to implement, but are typically numerically intensive.

The basis for all these methods is the idea that one can “simulate” the variance of a statistic by (i) making repeated draws from a distribution whose variance is related in a simple (and known) way to the variance of interest; (ii) empirically estimating the variance of this “secondary” distribution, and (iii) adjusting this variance estimate so that it is an estimate of the variance of interest.

2.4.2.1 Random groups estimate of variance

The simplest way of implementing the above idea is through the use of interpenetrating samples, see Mahalonobis (1946), Deming (1956). Here the actual sample selected is made up of *G* independent replicate or interpenetrating subsamples, each one of which is “representative” of the population, being drawn according to the same design and with the same sample size *n*/*G*. Let $\hat{\theta}_g$ denote the estimate of the *FPP* θ based on the g^{th} replicate sample. The overall estimate of this quantity is the average $\hat{\theta}$ of these $\hat{\theta}_g$.

By construction, the set of replicate estimates $\left\{\hat{\theta}_g, g=1,\ldots,G\right\}$ are independent and identically distributed. Consequently, we can estimate the variance of their (common) distribution by their empirical variance around their average, the overall estimate $\hat{\theta}$. Furthermore the variance of $\hat{\theta}$ is just this “replicate variance” divided by the number of replicates, *G*. Consequently we can estimate the variance of $\hat{\theta}$ by simply dividing this empirical variance by *G*, leading to the estimate

$$\hat{\mathrm{V}}_R\left(\hat{\theta}\right)=\frac{1}{G(G-1)}\sum_{g=1}^{G}\left(\hat{\theta}_g-\hat{\theta}\right)^2.$$

In fact, the above idea still works even if the replicate estimates are not identically distributed. All that is required is that they are independent of one another, and each is unbiased for the *FPP* θ. Straightforward algebra can then be used to show

$$\mathrm{E}\left(\hat{\mathrm{V}}_R\left(\hat{\theta}\right)\right)=\frac{1}{G^2}\sum_{g=1}^{G}\mathrm{V}\left(\hat{\theta}_g\right)=\mathrm{V}\left(\hat{\theta}\right)$$

so the replicate variance estimate is still unbiased for the variance of the average of the replicate estimates.

In practice replicated sample designs as described above are rare. However, the idea of replication-based variance estimation is still applicable. What is done in these cases is to construct the replicates after the sample is selected, by randomly allocating sample units to *G* groups in such a way that each group is at least approximately independent of the other groups.

With stratified designs such post-sample random grouping can be accomplished by random grouping within the strata, provided there is sufficient sample size within each stratum to carry this out. If this is not the case, then random grouping can be applied to the sample as a whole, preserving the strata when splitting the sample between the groups. In the case of multistage designs, splitting is typically carried out at *PSU* (primary sampling unit) level. In addition, the "average" estimate $\hat{\theta}$ in the variance formula above is often replaced by the "full sample" estimate of this quantity.

Finally, it should be pointed out that the replication variance estimate is an estimate of the variance (either design-based or model-based) of $\hat{\theta}$, not the variance of the sample error $\hat{\theta}-\theta$. A consequence is that this variance estimate does not go to zero as the sample size approaches the population size. This is of no great concern when sample sizes within strata are small compared to stratum population sizes. However, in many business surveys, sample sizes within strata are a substantial fraction of the stratum populations. In such cases, it is standard to multiply the stratum level replicated groups variance estimates by appropriate finite population correction factors.

2.4.2.2 Jackknife estimate of variance

A problem with the replication-based approach to variance estimation is the stability of these estimates. Clearly, the more groups there are, the more stable these variance estimates are. However, the more groups there are, the harder it is to "randomly group" the sample. A methodology that circumvents this problem, but at the cost of dropping the property of independent subgroup estimates, is to use overlapping groups.

There are essentially two approaches to using overlapping groups. The first is via Balanced Repeated Replication (*BRR*) where the groups are formed using experimental design precepts so that covariances induced by the same unit belonging to different groups "cancel out" in the (non-overlapping) random groups variance formula above. This can be quite difficult to accomplish in general, and so this method is typically restricted to certain types of multistage designs that are rarely used in business surveys. See Wolter (1985) and Shao & Tu (1995). The second, and more common method, is to compute a jackknife variance estimate.

Under the jackknife approach, the sample is again divided into G groups, but this time G estimates are computed by "dropping out" each of the G groups from the sample in turn. The variability between these dependent estimates is then used to estimate the variability of the overall estimate of θ. Let $\hat{\theta}_{(g)}$ denote the estimate of θ based on the sample excluding group g. The jackknife estimate of variance is

$$\hat{\text{V}}_J\left(\hat{\theta}\right)=\frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{\theta}_{(g)}-\hat{\theta}\right)^2 .$$

As with the replicated groups variance estimate, there are two forms of the jackknife variance estimate. The first, which we refer to as the *Type 1* jackknife, defines $\hat{\theta}$ as the average of the $\hat{\theta}_{(g)}$. The second, the *Type 2* jackknife, defines $\hat{\theta}$ as the "full sample" estimate of θ. Since

$$\sum_{g=1}^{G}\left(\hat{\theta}_{(g)}-\hat{\theta}\right)^2=\sum_{g=1}^{G}\left(\hat{\theta}_{(g)}-\frac{1}{n}\sum_{h=1}^{G}\hat{\theta}_{(h)}\right)^2+G\left(\hat{\theta}-\frac{1}{n}\sum_{h=1}^{G}\hat{\theta}_{(h)}\right)^2$$

the *Type 2* jackknife will be more conservative than the *Type 1* jackknife.

Unbiasedness of the jackknife variance estimate does not follow as easily as unbiasedness of the replicated groups variance estimate. For the *Type 1* jackknife, sufficient conditions for unbiasedness are

$$\text{V}\left(\hat{\theta}_{(g)}\right)=\frac{G}{G-1}\text{V}\left(\hat{\theta}\right)$$
$$\text{C}\left(\hat{\theta}_{(g)},\hat{\theta}_{(h)}\right)=\frac{G(G-2)}{(G-1)^2}\text{V}\left(\hat{\theta}\right).$$

For the *Type 2* jackknife the second condition above should be replaced by

$$\text{C}\left(\hat{\theta}_{(g)},\hat{\theta}\right)=\text{V}\left(\hat{\theta}\right).$$

As with the random groups variance estimate, the jackknife variance estimate is typically computed at *PSU* level in multistage samples. That is, the G groups are defined as groups of *PSU*s. Furthermore, the most common type of jackknife is when G is equal to the number of *PSU*s in sample, that is one *PSU* is dropped from the sample each time a value of $\hat{\theta}_{(g)}$ is calculated. There is empirical evidence that, provided the target parameter θ is sufficiently "smooth", this choice of G minimises the variance of the estimate of variance (Shao & Tu, 1995; example 2.1.4). Finally, one can note that, like the random groups variance estimate, the jackknife variance estimate does not include a finite population correction. This needs to be applied separately.

2.4.2.3 The linearised jackknife

The computational demands of the jackknife when $G = n$ (the number of sample *PSU*s) has led to research into ways of approximating it so that it can be computed in one "pass" of the sample data. If $\hat{\theta}$ is a "smooth" function of the sample data, this can be accomplished by essentially replacing $\hat{\mathrm{V}}_J(\hat{\theta})$ by a first order Taylor series approximation to it.

In what follows we assume single stage sampling. Furthermore, we assume the existence of a superpopulation model ξ under which $\mathrm{E}_\xi(y_j) = \mu_j$ for $j \in s$. Let μ denote the n-vector of these sample expected values. We can then approximate $\hat{\theta}$ by

$$\hat{\theta} = \hat{\theta}(\mu) + \sum_{j \in s} \left(\frac{\partial \hat{\theta}}{\partial y_j} \right)_\mu (y_j - \mu_j)$$

where $\hat{\theta}(\mu)$ denotes the value of $\hat{\theta}$ when the sample Y-values are replaced by μ and the partial derivatives in the second term on the right hand side are evaluated at μ as well. Similarly, let $\mu_{(j)}$ denote μ with the expected value for y_j deleted, and put $\hat{\theta}_{(j)}$ equal to the estimate based on the sample excluding y_j. The corresponding approximation to $\hat{\theta}_{(j)}$ is then

$$\hat{\theta}_{(j)} = \hat{\theta}_{(j)}(\mu_{(j)}) + \sum_{\substack{k \in s \\ k \neq j}} \left(\frac{\partial \hat{\theta}_{(j)}}{\partial y_k} \right)_{\mu_{(j)}} (y_k - \mu_k)$$

where $\hat{\theta}_{(j)}(\mu_{(j)})$ denotes $\hat{\theta}_{(j)}$ evaluated at $\mu_{(j)}$. We now make two extra assumptions:

(1) $\hat{\theta}(\mu) = \hat{\theta}_{(j)}(\mu_{(j)}) = \theta_0$;

(2) $\left(\frac{\partial \hat{\theta}}{\partial y_k} \right)_\mu = \frac{n}{n-1} \left(\frac{\partial \hat{\theta}_{(j)}}{\partial y_k} \right)_{\mu_{(j)}}$.

The first of these assumptions is uncontroversial, since it essentially corresponds to the requirement that the "drop out 1" and full sample estimates are estimating the same thing. The second assumption is reasonable when $\hat{\theta}$ is linear in Y, but may not be reasonable in other cases. With these assumptions we can replace the approximation to $\hat{\theta}_{(j)}$ above by

$$\hat{\theta}_{(j)} = \frac{n}{n-1} \left\{ \hat{\theta} - \left(\frac{\partial \theta}{\partial y_j} \right)_\mu (y_j - \mu_j) \right\} - \frac{\theta_0}{n-1}.$$

Substituting this approximation into the *Type 1* jackknife variance estimate leads to the linearised version of this estimate

$$\hat{\mathrm{V}}_{JL}^{(1)}(\hat{\theta}) = \frac{n}{n-1} \sum_{j \in s} \left\{ \left(\frac{\partial \hat{\theta}}{\partial y_j} \right)_{\hat{\mu}} (y_j - \hat{\mu}_j) - \frac{1}{n} \sum_{k \in s} \left(\frac{\partial \hat{\theta}}{\partial y_k} \right)_{\hat{\mu}} (y_k - \hat{\mu}_k) \right\}^2$$

where $\hat{\mu}$ denotes the full sample estimate of μ. The corresponding linearised *Type 2* jackknife is obtained similarly, after replacing θ_0 by $\hat{\theta}$. It is

$$\hat{\mathrm{V}}_{JL}^{(2)}(\hat{\theta}) = \frac{n}{n-1}\sum_{j \in s}\left\{\left(\frac{\partial \hat{\theta}}{\partial y_j}\right)_{\hat{\mu}}(y_j - \hat{\mu}_j) - \hat{\theta}\left(\frac{n^2 - 3n + 1}{n(n-1)}\right)\right\}^2 .$$

Comparing the preceding two expressions one can easily see that the linearised *Type 2* jackknife variance estimate will always be greater than the linearised *Type 1* jackknife variance estimate, a property that is generally observed for *Type 2* jackknife variance estimates.

Note that the linearised jackknife is essentially a model-based variance estimation procedure, since it requires specification and estimation of μ. Furthermore, it is unclear whether it leads to anything substantially different from using the Taylor approximation approach within a model-based framework for variance estimation. For example, the linearised *Type 1* jackknife estimate of the variance of the linear estimator $\hat{t}_L$ defined in 2.3.2.9,

$$\hat{\mathrm{V}}_{JL}^{(1)}(\hat{t}_L) = \frac{n}{n-1}\sum_{j \in s}\left\{w_{js}(y_j - \hat{\mu}_j) - \frac{1}{n}\sum_{k \in s} w_{ks}(y_k - \hat{\mu}_k)\right\}^2$$

is (to a first order approximation) equivalent to the robust model-based variance estimator $\hat{\mathrm{V}}_{\xi}^{R}(\hat{t}_L - t)$ described in 2.3.2.10.

2.4.2.4 Bootstrapping

Both the random groups and the jackknife methods result in estimates of variance for a statistic that is an estimate of a *FPP*. In general, however, our interest in such estimates is based on the desire to compute interval estimates (for example confidence intervals) for this *FPP*. Such quantities are defined in terms of the properties of the sampling distribution of the estimate. For large samples, the central limit theorem typically applies, and this sampling distribution can be well approximated by a normal distribution. In such cases it is sufficient (provided the estimate is asymptotically unbiased for the *FPP*) to estimate the variance of the sampling distribution in order to write down confidence intervals for this *FPP*.

However, for many sampling designs the level at which variances are calculated can be quite detailed (for example fine strata or domains containing relatively few units). Here an assumption of central limit behaviour may be quite inappropriate, in the sense that the sampling distribution (either design-based or model-based) may be quite non-normal. In these cases we may want to compute an estimate of the sampling distribution directly. The bootstrapping idea provides a way by which this objective can be achieved.

To start, we describe a model-based bootstrap, since this is relatively straightforward. In particular, we assume that the *FPP* of interest is defined in terms of the population values of a single *Y*-variable whose superpopulation distribution is specified by the model in 2.3.2.1, and a model-unbiased estimate $\hat{\omega}$ of the parameter ω in this model can be calculated from the sample data.

Let $\{r_{std,j}; j \in s\}$ denote the set of studentised residuals generated by the sample data under this model. That is, these residuals depend on $\hat{\omega}$ and satisfy $\mathrm{E}_\xi(r_{std,j}) = 0$ and $\mathrm{V}_\xi(r_{std,j}) = 1$. By sampling at random with replacement from $\{r_{std,j}; j \in s\}$ we can then generate a set of N bootstrap residuals $\{r_k^*; k \in U\}$ and consequently a bootstrap realisation of the population values of Y, defined by

$$y_k^* = \mu(x_k;\hat{\omega}) + \sigma(x_k;\hat{\omega})r_k^*.$$

Given this bootstrap realisation, we can compute a bootstrap estimate of θ based on the values $\{y_j^*; j \in s\}$, which we denote by $\hat{\theta}^*$, together with the actual value of θ for the bootstrap population, which we denote by θ^*. The bootstrap realisation of the sample error is then $\hat{\theta}^* - \theta^*$. This process is now repeated a large number of times, leading to a distribution of such bootstrap sample errors. We denote the mean of this bootstrap distribution by $\mathrm{E}^*(\hat{\theta}^* - \theta^*)$, and its variance by $\mathrm{V}^*(\hat{\theta}^* - \theta^*)$.

The bootstrap estimate of θ is then $\hat{\theta}_B = \hat{\theta} + \mathrm{E}^*(\hat{\theta}^* - \theta^*)$. The bootstrap variance of this estimate is sometimes taken as $\mathrm{V}^*(\hat{\theta}^* - \theta^*)$. However, this will typically be an underestimate since it does not take account of the error in estimation of ω in the above process. Consequently it is usually better to rescale the bootstrap sample error distribution so that its variance is the larger of this initial variance or an estimate of the variance which allows for error in estimation of ω (for example, a jackknife estimate). If it is also believed that $\hat{\theta}$ represents a "best" estimate of θ, then the bootstrap sample error distribution can be centred at zero prior to this rescaling.

In any case, after recentering and rescaling, it is simple to "read off" a $100(1-\alpha)\%$ confidence interval for θ from the bootstrap sample error distribution. Essentially such a confidence interval is defined by

$$\left(\hat{\theta}_B - Q^*\left(\frac{\alpha}{2}\right), \hat{\theta}_B + Q^*\left(1 - \frac{\alpha}{2}\right)\right)$$

where $Q^*(\gamma)$ denotes the γ-th quantile of this distribution.

One problem with the bootstrap procedure defined above is that it depends on correct specification of the heteroskedasticity function $\sigma(x; \omega)$. A heteroskedasticity-robust model-based bootstrap is easily defined, however. Essentially, all one needs to do is to replace the studentised residuals underpinning the bootstrap procedure by "raw" residuals $r_{raw,j} = y_j - \mu(x_j;\hat{\omega})$. The remaining steps in the bootstrap procedure are unchanged. See Chambers & Dorfman (1994).

Bootstrapping the design-based distribution of the sample error is also possible, but can be quite complicated depending on the actual survey design used. This is because one needs to sample with replacement from the sample Y-values in such a way as to at least "preserve" the first and second order inclusion probabilities of the design. Consequently, at the time of writing, a number of "bootstrap-type" methods for estimating the design variance have been suggested (Shao & Tu, 1995, Chapter 6), with no obvious preferred method.

The simplest of these at present is the bootstrap procedure described by Canty & Davison (1997). We describe this in the context of estimation of the variance of the linear estimate $\hat{t}_L$ defined in 2.3.2.9, where the sample weights are calibrated to the population total of an auxiliary variable X. That is, when the estimate $\hat{t}_L$ is calculated with the sample Y-values replaced by sample X-values, the known population total of X is obtained. A bootstrap replication here consists of the following steps:

(1) select a simple random sample of n labels from s with replacement. Let i index the n draws making up this bootstrap sample. Thus y_i^* denotes the value of Y corresponding to the sample label selected at the i^{th} draw, w_{is}^* denotes the sample weight associated with this value, and x_i^* denotes the corresponding value of the auxiliary variable;

(2) recalibrate the weights associated with the bootstrap sample. Let w_i^* denote the recalibrated weight associated with the i^{th} bootstrap sample Y-value;

(3) recompute the bootstrap realisation of $\hat{t}_L$. Assuming $\hat{t}_L$ is a *GREG* estimate, this will be of the form:

$$\hat{t}_L^* = \sum_{i=1}^{n} w_i^* y_i = \sum_{i=1}^{n} w_{is}^* y_i^* + \hat{\beta}^* \left(\sum_{j \in s} w_{js} x_j - \sum_{i=1}^{n} w_{is}^* x_i^* \right)$$

where $\hat{\beta}^*$ denotes the estimate of the regression of Y on X based on the bootstrap sample.

Repeating the above procedure a large number of times then generates the bootstrap distribution of $\hat{t}_L$. As usual we denote the mean and variance of this bootstrap distribution (that is, conditional on the sample Y-values) by E* and V* respectively. The bootstrap variance estimate is the empirical variance of the bootstrap values $\hat{t}_L^*$ over these replications.

Although exact expressions for the moments of the above bootstrap distribution are generally unavailable, good approximations are easily worked out. For any particular bootstrap replication, define I_{ji}^* as one if the j^{th} sample unit was selected at the i^{th} draw making up the bootstrap sample selected at that replication, and as zero otherwise. Then

$$I_j^* = \sum_{i=1}^{n} I_{ji}^*$$

denotes the number of times the j^{th} sample unit contributes to this bootstrap sample. It follows $\mathrm{E}^*\left(I_j^*\right)=1$, $\mathrm{V}^*\left(I_j^*\right)=(n-1)/n$ and $\mathrm{C}^*\left(I_j^*, I_k^*\right)=-1/n$. Furthermore, since we can write

$$\hat{t}_L^* = \sum_{j \in s} w_{js} y_j I_j^* + \hat{\beta}^* \left(\sum_{j \in s} w_{js} x_j - \sum_{j \in s} w_{js} x_j I_j^* \right)$$

we can approximate this bootstrap realisation of $\hat{t}_L$ by replacing $\hat{\beta}^*$ by the coefficient $\hat{\beta}$ of the "full sample" regression of *Y* on *X*. With this approximation it is easy to see that $\mathrm{E}^*\left(\hat{t}_L^*\right) = \hat{t}_L$, while

$$\begin{aligned}
\mathrm{V}^*\left(\hat{t}_L^*\right) &= \mathrm{V}^*\left(\sum_{j \in s} w_{js}\left(y_j - \hat{\beta}x_j\right) I_j^* \right) \\
&= \sum_{j \in s} w_{js}^2 \left(y_j - \hat{\beta}x_j\right)^2 \mathrm{V}^*\left(I_j^*\right) + \sum_{j \in s} \sum_{\substack{k \in s \\ k \neq j}} w_{js} w_{ks} \left(y_j - \hat{\beta}x_j\right)\left(y_k - \hat{\beta}x_k\right) \mathrm{C}^*\left(I_j^*, I_k^*\right) \\
&= \sum_{j \in s} w_{js}^2 \left(y_j - \hat{\beta}x_j\right)^2 - \frac{1}{n}\left(\sum_{j \in s} w_{js}\left(y_j - \hat{\beta}x_j\right) \right)^2 \\
&= \frac{n-1}{n} \hat{\mathrm{V}}_{JL}^{(1)}\left(\hat{t}_L\right).
\end{aligned}$$

That is, this first order approximation to the Canty-Davison bootstrap variance estimate is $(n - 1)/n$ times the linearised *Type 1* jackknife variance estimate. Clearly, this approximation is exactly the jackknife variance estimate provided we modify the bootstrap procedure above to select $n - 1$ rather than n sample labels at each replication.

2.5 Conclusions

The purpose of this chapter has been to set out the basic theory for sampling error related bias and variance assessment of standard survey estimates. This theory has either depended on, or required, the use of some form of probability sampling method. Two basic paradigms for defining bias and variance have been presented: the design-based approach which measures these quantities relative to the uncertainty associated with the different samples that could have been selected under the method used; and the model-based approach which measures the uncertainty in terms of the possible values that the survey variable can take in the target population. Both approaches have strengths and weaknesses, and these have been pointed out. In the end, it seems clear that robust model-based/model-assisted methods and sensibly conditioned design-based methods for assessing bias and variance tend to lead to similar conclusions, and so this chapter has attempted, where possible, to indicate the connection between the two.

From the point of view of best practice as far as minimisation of sampling bias and assessment of sampling variance are concerned, we suggest the following points be kept in mind:

- robust probability sampling methods should be used wherever possible. These are designs which blend randomisation and modelling ideas in order to ensure that the samples that are finally selected are not only "random" but also representative of the full range of potential *Y*-values under a carefully specified model for the target population. Such samples are necessary if the size of the sampling error is to be kept within acceptable bounds;

- robust methods of sampling variance estimation should be used if at all possible. Given the representative "balanced" samples that arise under the preceding recommendation, these methods provide stable and accurate assessments of the potential size of the sample error. However, it should also be kept in mind that these methods are not guaranteed to work if the sample is unrepresentative. Essentially all robust methods for estimating sample error variability assume that the variability in the sample values is representative of that in the target population. This is not the case if the sample is unrepresentative;
- for complex *FPP*'s one has a choice between "plug-in" methods based on Taylor series linearisation arguments or a variety of replication or resampling methods. The former are less computer intensive but (sometimes) require considerable analytic skill to develop and program. The latter are generally easy to program but are typically highly computer intensive. The choice between these methods depends on the resources at hand. Some appreciation for the different operating characteristics of these methods can be obtained by reading the volume of this report dealing with assessment of different computer software for survey inference. It suffices to point out that generally, because of their "plug-in" nature, Taylor series linearisation methods tend to underestimate sampling variability, while replication/resampling methods tend to overestimate it. In medium to large samples, however, there is little to choose between these methods since all are essentially first order equivalent.

3 Probability sampling: extensions

Ray Chambers, University of Southampton

3.1 Domain estimation

A common problem in survey inference is estimation of the population total of a survey variable *Y* for a domain of interest. For example, in many business surveys the sample frame is out of date, so the industry and size classifications of many units on the frame do not agree with their "current" industry and size classifications. After the survey is carried out, estimates are required for the current industry by size classes. These classes then correspond to domains of interest as far as the survey is concerned.

In general, a domain is some subgroup of the sample population. Often domains cut across stratum boundaries and are referred to as "cross-classes". *A basic assumption in domain estimation is that domain membership is observable on the sample*. That is, one can define a domain membership variable *D* with value d_j for population unit *j*, such that $d_j = 1$ if unit *j* is in the domain and is zero otherwise, and the values of *D* are observable for the sample units. The number of population units in the domain is just the population sum of *D* and is denoted by N_d. By construction, the population total for the domain is

$$t_d = \sum_{j \in U} d_j y_j \ .$$

3.1.1 Design-based inference for domains

Within the design-based framework, domain estimation poses no special problems. It is sufficient to note that the domain total is just the population total of the variable *DY*. Consequently the *HTE* for t_d is just

$$\hat{t}_{dHT} = \sum_{j \in s} \pi_j^{-1} d_j y_j$$

with design variance

$$\mathrm{V}_p\left(\hat{t}_{dHT} - t_d\right) = \sum_{j \in U}\sum_{k \in U} \frac{\left(\pi_{jk} - \pi_j \pi_k\right) d_j y_j d_k y_k}{\pi_j \pi_k}$$

The *SYG* estimate of the variance of this estimate is

$$\hat{\mathrm{V}}_p^{SYG}\left(\hat{t}_{dHT} - t_d\right) = \frac{1}{2} \sum_{j \in s} \sum_{\substack{k \in s \\ k \neq j}} \left(\frac{\pi_j \pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(\frac{d_j y_j}{\pi_j} - \frac{d_k y_k}{\pi_k}\right)^2 .$$

3.1.2 Design-based inference under SRSWOR

The case of simple random sampling without replacement (SRSWOR) is instructive, since it is the one situation where model-based inference and design-based inference "come together". In this case

$$\hat{t}_{dHT} = \frac{N}{n}\sum_{j \in s} d_j y_j = N p_{sd} \bar{y}_{sd}$$

where $\bar{y}_{sd}$ is the sample average of Y for units in the domain, and p_{sd} is the sample proportion of units in the domain. This estimator is intuitively reasonable. One modifies an estimate of the population total that effectively treats all population units as belonging to the domain by an estimate of the proportion of population units that actually belong to the domain. The design variance of this estimator is (after some algebra)

$$\mathrm{V}_p\left(\hat{t}_{dHT} - t_d\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left[p_d s_d^2 + p_d\left(1 - p_d\right)\bar{y}_d^2\right]$$

where $p_d = N_d / N$ is the proportion of population units that are in the domain, $\bar{y}_d$ is the average value of Y in the domain and s_d is the standard deviation of the Y-values in the domain. Ignoring $O\left(N_d^{-1}\right)$ terms, the SYG estimate of this variance is

$$\hat{\mathrm{V}}_p^{SYG}\left(\hat{t}_{dHT} - t_d\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left[p_{sd} s_{sd}^2 + p_{sd}\left(1 - p_{sd}\right)\bar{y}_{sd}^2\right]$$

where s_{sd} is the standard deviation of the sample Y-values in the domain.

3.1.3 Model-based inference when *N*

Model-based inference for a domain total depends on what one knows about the domain, and in particular on whether one knows how many population units are in the domain. That is, it depends on whether one knows the value of N_d. It also depends on whether the method of sample selection depends on domain inclusion or not (remember we are assuming that the sampling method is uninformative as far as Y is concerned). To start, we consider the most common situation, where the value of N_d is unknown.

To illustrate the model-based approach, consider the case where the estimator of choice is the *HTE* defined in 3.1.2. As usual, we let a subscript of ξ denote quantities defined with respect to a superpopulation model ξ. The particular model we assume is very simple and is specified by

$$\begin{aligned} \mathrm{E}_\xi\left(y_j \mid d_j = 1\right) &= \mu_d & \mathrm{E}_\xi\left(d_j\right) &= \theta_d \\ \mathrm{V}_\xi\left(y_j \mid d_j = 1\right) &= \sigma_d^2 & \mathrm{V}_\xi\left(d_j\right) &= \theta_d\left(1 - \theta_d\right) \\ \mathrm{C}_\xi\left(y_j, y_k \mid d_j, d_k\right) &= 0 & \mathrm{C}_\xi\left(d_j, d_k\right) &= 0 \end{aligned}$$

That is, domain membership in the population is modelled as the outcome of a Bernoulli process with fixed "success" probability θ_d, and conditional on domain membership the population values of Y are uncorrelated with constant mean and variance.

As with the model-based approach in general, there is an implicit assumption that sample inclusion is independent of the values of the variables of interest. In this context, this requires that sample inclusion and domain membership be independent of one another. This assumption is valid if the sample is chosen via simple random sampling.

Under the above model it is easy to see that

$$\begin{aligned} \mathrm{E}_{\xi}\left(d_j y_j\right) &= \mu_d \theta_d \\ \mathrm{V}_{\xi}\left(d_j y_j\right) &= \sigma_d^2 \theta_d + \mu_d^2 \theta_d \left(1-\theta_d\right) \\ \mathrm{C}_{\xi}\left(d_j y_j, d_k y_k\right) &= 0 \end{aligned}$$

so the Best Linear Unbiased Predictor (*BLUP*) for t_d is just the *HTE*. Furthermore the model variance of the *HTE*/*BLUP* is

$$\mathrm{V}_{\xi}\left(\hat{t}_{dHT} - t_d\right) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left[\theta_d \sigma_d^2 + \theta_d \left(1-\theta_d\right)\mu_d^2\right]$$

so the *SYG* variance estimate in 3.1.2 is also an unbiased estimate of this model variance. For this case, model-based and design-based inference coincide.

3.1.4 Model-based inference when N_d is known

Here one is lead to inference that *conditions* on this known value of N_d. To illustrate, we consider the same situation as in 3.1.3. In this case, however, we need to modify the model considered there to take account of the extra information provided by knowledge of N_d. Let $\mathrm{E}_{\xi d}, \mathrm{V}_{\xi d}, \mathrm{C}_{\xi d}$ denote expectation, variance and covariance conditional on knowing N_d. As before we put $p_d = N_d / N$. Then, since $\mathrm{E}_{\xi d}\left(N_d\right) = N_d$ and $\mathrm{V}_{\xi d}\left(N_d\right) = 0$, symmetry-based arguments can be used to show that

$$\begin{aligned} \mathrm{E}_{\xi d}\left(d_j\right) &= p_d \\ \mathrm{V}_{\xi d}\left(d_j\right) &= p_d \left(1-p_d\right) \\ \mathrm{C}_{\xi d}\left(d_j, d_k\right) &= -p_d \left(1-p_d\right) / (N-1) \end{aligned}$$

Furthermore, if we assume that *Y* is independent of N_d conditional on *D* (that is, knowing N_d tells us nothing extra about y_j than knowing the value of d_j), and the conditional moments of *Y* given *D* are as specified in 3.1.3, then the following results hold

$$\begin{aligned} \mathrm{E}_{\xi d}\left(d_j y_j\right) &= \mu_d p_d \\ \mathrm{V}_{\xi d}\left(d_j y_j\right) &= \sigma_d^2 p_d + \mu_d^2 p_d \left(1-p_d\right) \\ \mathrm{C}_{\xi d}\left(d_j y_j, d_k y_k\right) &= -\mu_d^2 p_d \left(1-p_d\right) / (N-1) \\ \mathrm{C}_{\xi d}\left(d_j y_j, d_j\right) &= -\mu_d p_d \left(1-p_d\right) \\ \mathrm{C}_{\xi d}\left(d_j y_j, d_k\right) &= -\mu_d p_d \left(1-p_d\right) / (N-1) \end{aligned}$$

From the first three identities above we see that, with respect to this conditional distribution, the “derived” random variable *DY* has a mean and variance that is the same for all population units. Furthermore, the covariance between any two population values of *DY* is constant. It is straightforward to show that the *BLUP* defined in terms of this "derived" variable is then still the *HTE*. In fact, we have

$$\begin{aligned}\mathrm{V}_{\xi d}\left(\hat{t}_{dHT}-t_d\right)&=\frac{N^2}{n}\left(1-\frac{n}{N}\right)\left[\mathrm{V}_{\xi d}\left(d_j y_j\right)-\mathrm{C}_{\xi d}\left(d_j y_j, d_k y_k\right)\right]\\&=\frac{N^2}{n}\left(1-\frac{n}{N}\right)\left[\sigma_d^2 p_d+\frac{N}{N-1}\mu_d^2 p_d\left(1-p_d\right)\right].\end{aligned}$$

However, in this situation there seems no strong reason why one should restrict attention to estimates that are linear in *DY*. An obvious alternative is the nonlinear ratio-type estimate

$$\hat{t}_{dR}=N_d\bar{y}_{sd}=N_d\frac{\sum_{j\in s}d_j y_j}{\sum_{j\in s}d_j}$$

This estimate is approximately model-unbiased in large samples. Furthermore, the variance of this estimate can be approximated using a standard Taylor series argument. In fact, one can show

$$\begin{aligned}\mathrm{V}_{\xi d}\left(\hat{t}_{dR}-t_d\right)&\approx\frac{N^2}{n^2}\mathrm{V}_{\xi d}\left[\sum_{j\in s}d_j y_j-\mu_d\sum_{j\in s}d_j-\frac{n}{N}\sum_{j\in U}d_j y_j\right]\\&=\frac{N^2}{n^2}\left(1-\frac{n}{N}\right)\sigma_d^2 p_d.\end{aligned}$$

Comparing this variance with the variance of the *HTE*, we see that there will typically be large efficiency gains from use of the ratio-type estimate.

There is a fundamental principle sometimes invoked in model-based inference called the *conditionality principle* (Cox & Hinkley, 1974). This states that one should always condition on ancillary variables in inference. An ancillary variable is one whose distribution depends on parameters that are distinct from those associated with the distribution of the variable of interest. In the context of domain analysis, it can be argued that the parameter(s) associated with the distribution of the domain inclusion variable *D* are distinct from those associated with the distribution of the survey variable *Y*. Consequently, one should condition on *D* in inference. This is equivalent to conditioning on both the population count N_d of the number of units in the domain, *and* the corresponding sample count n_d.

If one conditions in this way it is straightforward to show that the ratio-type estimate above is the *BLUP* for t_d (defined in terms of *Y*) and has model variance

$$\mathrm{V}_{\xi}\left(\hat{t}_{dR}-t_d\mid n_d,N_d\right)=\frac{N_d^2}{n_d}\left(1-\frac{n_d}{N_d}\right)\sigma_d^2$$

This is sometimes referred to as the variance of the poststratified estimate for the domain total.

Which of the two immediately preceding variances for the ratio-type estimate is "correct" is the subject of debate. Clearly, "plug in" estimates for both will be different in general, with equality only if the population sampling fraction equals the domain sampling fraction. An argument against the poststratified approach is based on the fact that the distribution of the population parameter t_d depends on the parameters of Y as well as the parameters of D. Consequently this is a case where the ancillarity principle is not applicable. Raised against this, however, is the argument that, unlike the conditional variance, the poststratified variance is zero if $N_d = n_d$, when we *know* that the ratio-type estimate has zero error. However, often one will have $N_d >> n_d$ and so a cautious approach would be to estimate the variance of the ratio-type estimate by the maximum of the two variance estimates.

3.1.5 Model-based inference utilising auxiliary information

We return to the case where the domain count N_d is unknown. However, we extend the model for Y to the one considered in 2.3.2.1. That is, we assume

$$\begin{aligned}\mathrm{E}_\xi(y_j \mid d_j = 1) &= \mu(x_j;\omega_d)\\ \mathrm{V}_\xi(y_j \mid d_j = 1) &= \sigma^2(x_j;\omega_d)\\ \mathrm{C}_\xi(y_j, y_k \mid d_j, d_k) &= 0 \quad \text{for } j \neq k.\end{aligned}$$

We continue to assume that domain membership is defined by a sequence of independent and identically distributed Bernoulli trials, independently of the value of Y. However, domain membership *can* depend on X, so

$$\begin{aligned}\mathrm{E}_\xi(d_j) &= \theta(x_j;\gamma_d)\\ \mathrm{V}_\xi(d_j) &= \theta(x_j;\gamma_d)[1-\theta(x_j;\gamma_d)]\\ \mathrm{C}_\xi(d_j, d_k) &= 0.\end{aligned}$$

With this set-up we have

$$\begin{aligned}\mathrm{E}_\xi(d_j y_j) &= \mu(x_j;\omega_d)\theta(x_j;\gamma_d)\\ \mathrm{V}_\xi(d_j y_j) &= \sigma^2(x_j;\omega_d)\theta(x_j;\gamma_d) + \mu^2(x_j;\omega_d)\theta(x_j;\gamma_d)[1-\theta(x_j;\gamma_d)]\\ \mathrm{C}_\xi(d_j y_j, d_k y_k) &= 0.\end{aligned}$$

Given probability sampling, consistent estimates for the parameters ω_d and γ_d above can be obtained from the sample data. A plug in model-based estimate of t_d is then

$$\hat{t}_{d\xi} = \sum_{j\in s} d_j y_j + \sum_{j\notin s} \mu(x_j;\hat{\omega}_d)\theta(x_j;\hat{\gamma}_d)$$

where a "hat" denotes a sample estimate. Clearly this estimate will also be consistent.

The model-variance of this estimate can be written

$$\mathrm{V}_\xi(\hat{t}_{d\xi} - t_d) = \mathrm{V}_\xi\left(\sum_{j\notin s} \mu(x_j;\hat{\omega}_d)\theta(x_j;\hat{\gamma}_d)\right) + \sum_{j\notin s} \mathrm{V}_\xi(d_j y_j) = \mathrm{V}_{1\xi} + \mathrm{V}_{2\xi}$$

The leading (biggest) term in this variance is V$_{1\xi}$. It can be estimated using computer intensive methods like the jackknife or bootstrap. For example, the "drop-out 1" Type 2 jackknife estimate of this quantity is

$$\hat{\mathrm{V}}_{J\xi}^{(1)} = \frac{n-1}{n}\sum_{j\in s}\left(\sum_{k\notin s}\mu\left(x_k;\hat{\omega}_{d(j)}\right)\theta\left(x_k;\hat{\gamma}_{d(j)}\right) - \sum_{k\notin s}\mu\left(x_k;\hat{\omega}_d\right)\theta\left(x_k;\hat{\gamma}_d\right)\right)^2$$

where $\hat{\omega}_{d(j)}$ denotes the sample estimate of ω_d based on the sample units excluding unit *j*, and $\hat{\gamma}_{d(j)}$ is defined similarly. Typically $\hat{\omega}_{d(j)}$ is just $\hat{\omega}_d$ for all sample units not in the domain, so some simplification of the above formula is possible.

Alternatively, a Taylor series linearisation approach can be used to construct a "direct" estimate of V$_{1\xi}$. This is based on the approximation

$$\mathrm{V}_{1\xi} \approx \mathrm{V}_\xi\left(\hat{\gamma}_d\sum_{j\in s}\mu\left(x_j;\hat{\omega}_{0d}\right)\frac{\partial\theta\left(x_j;\gamma_{0d}\right)}{\partial\gamma_{0d}} + \hat{\omega}_d\sum_{j\notin s}\theta\left(x_j;\gamma_{0d}\right)\frac{\partial\mu\left(x_j;\omega_{0d}\right)}{\partial\omega_{0d}}\right)$$

where γ_{0d} and ω_{0d} are the "true" values of γ_0 and ω_0, and the partial derivatives are evaluated at these "true" values.

Depending on the specification of the functions μ and θ, estimates of the variances of $\hat{\omega}_d$ and $\hat{\gamma}_d$ and their covariance can be estimated from the sample data. Using "hats" to denote these estimates in the usual way, this suggests a Taylor series estimate of V$_{1\xi}$ of the form

$$\begin{aligned}\hat{\mathrm{V}}_{1\xi} = {} & \hat{\mathrm{V}}_\xi(\hat{\gamma}_d)\left(\sum_{j\notin s}\mu\left(x_j;\hat{\omega}_d\right)\frac{\partial\theta\left(x_j;\hat{\gamma}_d\right)}{\partial\hat{\gamma}_d}\right)^2 + \hat{\mathrm{V}}_\xi(\hat{\omega}_d)\left(\sum_{j\notin s}\theta\left(x_j;\hat{\gamma}_d\right)\frac{\partial\mu\left(x_j;\hat{\omega}_d\right)}{\partial\hat{\omega}_d}\right)^2 \\ & + 2\hat{\mathrm{C}}_\xi(\hat{\gamma}_d,\hat{\omega}_d)\left(\sum_{j\notin s}\mu\left(x_j;\hat{\omega}_d\right)\frac{\partial\theta\left(x_j;\hat{\gamma}_d\right)}{\partial\hat{\gamma}_d}\right)\left(\sum_{j\notin s}\theta\left(x_j;\hat{\gamma}_d\right)\frac{\partial\mu\left(x_j;\hat{\omega}_d\right)}{\partial\hat{\omega}_d}\right)\end{aligned}$$

The second term V$_{2\xi}$ in the variance formula has a simple plug-in estimate based on the model specification above. This is

$$\hat{\mathrm{V}}_{2\xi} = \sum_{j\notin s}\left(\sigma^2\left(x_j;\hat{\omega}_d\right)\theta\left(x_j;\hat{\gamma}_d\right) + \mu^2\left(x_j;\hat{\omega}_d\right)\theta\left(x_j;\hat{\gamma}_d\right)\left[1-\theta\left(x_j;\hat{\gamma}_d\right)\right]\right).$$

3.1.6 An example

A simple example illustrating the above theory is where the population is stratified and the regression of *Y* on *X* is linear and through the origin for units in the domain, but the slope of this regression line varies from stratum to stratum. Furthermore, the proportion of the population in the domain varies significantly from stratum to stratum. Here we put θ_h equal to the probability that a population unit in stratum *h* lies in the domain and β_h equal to the slope of the regression line for domain units in stratum h. Our estimate of the domain total of Y for the population is then

$$\hat{t}_{d\xi} = \sum_{j \in s} d_j y_j + \sum_h p_{shd}\left(N_h \bar{x}_h - n_h \bar{x}_{sh}\right)\hat{\beta}_h$$

Here h indexes the strata, p_{shd} denotes the sample proportion of stratum h units in the domain, $\hat{\beta}_h$ denotes the stratum h estimate for the slope of the regression of Y on X in the domain, $\bar{x}_h$ denotes the stratum average for X and $\bar{x}_{sh}$ is the sample average for X in stratum h. The Taylor series estimate of the leading term in the model-variance of this estimate is

$$\hat{\mathrm{V}}_{1\xi} = \sum_h \left(N_h \bar{x}_h - n_h \bar{x}_{sh}\right)^2 \left(\hat{\mathrm{V}}_\xi\left(p_{shd}\right)\hat{\beta}_h^2 + \hat{\mathrm{V}}_\xi\left(\hat{\beta}_h\right)p_{shd}^2\right)$$

where $\hat{\mathrm{V}}_\xi\left(p_{shd}\right)$ denotes the estimated variance of p_{hd} and $\hat{\mathrm{V}}_\xi\left(\hat{\beta}_h\right)$ denotes the estimated variance of $\hat{\beta}_h$. Note that independence of D and Y within a stratum causes the covariance term in this estimate to disappear. Typically

$$\hat{\mathrm{V}}_\xi\left(p_{shd}\right) = n_h^{-1} p_{shd}\left(1 - p_{shd}\right)$$

and, if we also assume that the residual variance for the regression of Y on X is proportional to X within a stratum by domain “cell”, then

$$\hat{\mathrm{V}}_\xi\left(\hat{\beta}_h\right) = \left(n_{hd}\bar{x}_{shd}\right)^{-1}\hat{\sigma}_h^2$$

where $\hat{\sigma}_h$ is the usual estimate of the residual scale parameter for this regression, n_{hd} is the number of domain units in sample in stratum h and $\bar{x}_{shd}$ is their average X-value. Substituting theses estimates and adding on $\hat{\mathrm{V}}_{2\xi}$ for this case leads to a variance estimate of the form

$$\begin{aligned}\hat{\mathrm{V}}_\xi = \hat{\mathrm{V}}_{1\xi} + \hat{\mathrm{V}}_{2\xi} = \sum_h p_{shd}\Bigg\{&\left(1 - p_{shd}\right)\hat{\beta}_h^2\left[\frac{\left(N_h \bar{x}_h - n_h \bar{x}_{sh}\right)^2}{n_h} + N_h \bar{x}_h^2 - n_h \bar{x}_{sh}^2\right] \\ &+ \hat{\sigma}_h^2\left[\frac{\left(N_h \bar{x}_h - n_h \bar{x}_{sh}\right)^2}{n_{hd}\bar{x}_{shd}} p_{shd} + \left(N_h \bar{x}_h - n_h \bar{x}_{sh}\right)\right]\Bigg\}\end{aligned}$$

where $\bar{x}_h^2$ denotes the average of X^2 in stratum h, and $\bar{x}_{sh}^2$ is the corresponding sample quantity. In the special case where $X \equiv 1$ it is straightforward to see that this expression reduces to the stratified random sampling version of the SYG variance estimate described in 3.1.2.

3.1.7 Domain estimation using a linear weighted estimate

Most computing packages for survey estimation which use a linear estimate of the form $\hat{t}_L$ described in 2.3.2.9 carry out domain estimation by simply replacing the y_j in this estimate by $d_j y_j$. That is, they calculate the linear weighted domain estimate

$$\hat{t}_{dL} = \sum_{j \in s} w_{js} d_j y_j \, .$$

Under the general domain model of 3.1.5 the model-bias of this estimate is

$$\mathrm{E}_\xi\left(\hat{t}_{dL} - t_d\right) = \sum_{j \in s} w_{js} \mu\left(x_j; \omega_d\right) \theta\left(x_j; \gamma_d\right) - \sum_{j \in U} \mu\left(x_j; \omega_d\right) \theta\left(x_j; \gamma_d\right).$$

There is no particular reason for this model-bias to be zero, or even close to zero. To illustrate, suppose (as is often the case) that the regression of *Y* on *X* in the population is linear in *X* and the weights w_{js} are calibrated on *X*. This is sufficient to ensure model-unbiasedness of $\hat{t}_L$. Suppose also that the regression of *Y* on *X* in the domain is linear in *X*, but with the addition of a domain "shift" term. That is

$$\mu\left(x_j; \omega_d\right) = x_j^{\mathrm{T}} \beta + \eta_d \, .$$

Then

$$\mathrm{E}_\xi\left(\hat{t}_{dL} - t_d\right) = \left(\sum_{j \in s} w_{js} x_j^{\mathrm{T}} \theta\left(x_j; \gamma_d\right) - \sum_{j \in U} x_j^{\mathrm{T}} \theta\left(x_j; \gamma_d\right) \right) \beta + \eta_d \left(\sum_{j \in s} w_{js} \theta\left(x_j; \gamma_d\right) - \sum_{j \in U} \theta\left(x_j; \gamma_d\right) \right).$$

Unless the domain inclusion probability does not depend on *X*, it is clear that both terms in this bias will be nonzero in general, irrespective of the calibrated nature of the weights.

One situation where the second term in the above bias disappears is where *X* includes stratum indicators, so the calibrated weights sum to the stratum population count within a stratum, and where domain inclusion probabilities are constant within a stratum. In this case (s_h denotes the stratum subsample, U_h denotes the stratum population)

$$\mathrm{E}_\xi\left(\hat{t}_{dL} - t_d\right) = \sum_h \theta_h \left(\sum_{j \in s_h} w_{js} z_j^{\mathrm{T}} - \sum_{j \in U_h} z_j^{\mathrm{T}} \right) \beta_z$$

where z_j denotes x_j with stratum indicators removed, and β_z is the corresponding component of β. Clearly this remaining model-bias will vanish if the weights are actually calibrated on *X* within strata, which is equivalent to requiring model-unbiasedness for $\hat{t}_L$ in the case where the linear regression model for *Y* includes interactions between the stratum indicator components of *X* and the remaining components of this auxiliary variable.

In principle, one can estimate the model-bias of the linear weighted domain estimate via

$$\hat{B}_\xi\left(\hat{t}_{dL}\right) = \sum_{j \in s} w_{js} \mu\left(x_j; \hat{\omega}_d\right) \theta\left(x_j; \hat{\gamma}_d\right) - \sum_{j \in U} \mu\left(x_j; \hat{\omega}_d\right) \theta\left(x_j; \hat{\gamma}_d\right)$$

and hence "correct" this estimate for its model-bias. For example, in the stratified case discussed above this bias estimate is

$$\hat{B}_\xi\left(\hat{t}_{dL}\right) = \sum_h N_h p_{shd} \left(\bar{z}_{wsh}^{\mathrm{T}} - \bar{z}_h^{\mathrm{T}} \right) \hat{\beta}_z$$

where $\bar{z}_{wsh}$ is the weighted average of the sample z_j in stratum h, and $\bar{z}_h$ is the actual stratum average for this auxiliary variable. The statistical properties of this bias corrected estimate are unknown at the time of writing.

3.1.8 Model-assisted domain inference

We focus on the extension of the GREG idea to domain estimation. The corresponding modification to the GRAT idea is straightforward. Thus, applying the GREG idea under the general model of 3.1.5 leads to the estimate

$$\tilde{t}_{dGREG} = \sum_{j \in U} \mu(x_j; \hat{\omega}_{pd}) \theta(x_j; \hat{\gamma}_{pd}) + \sum_{j \in s} \frac{d_j y_j - \mu(x_j; \hat{\omega}_{pd}) \theta(x_j; \hat{\gamma}_{pd})}{\pi_j}$$

where $\hat{\omega}_{pd}$ is a design consistent estimate of ω_d, and $\hat{\gamma}_{pd}$ is defined similarly. Defining residuals $\hat{e}_{dj} = d_j y_j - \mu(x_j; \hat{\omega}_{pd}) \theta(x_j; \hat{\gamma}_{pd})$, a first order approximation to the SYG estimate of the leading term in the design variance of this estimate is then easily seen to be

$$\hat{\mathrm{V}}_{\mathrm{p}}(\tilde{t}_{dGREG}) = \frac{1}{2} \sum_{j \in s} \sum_{\substack{k \in s \\ k \neq j}} \left(\frac{\pi_j \pi_k - \pi_{jk}}{\pi_{jk}} \right) \left(\frac{\hat{\theta}_{dj}}{\pi_j} - \frac{\hat{\theta}_{dk}}{\pi_k} \right)^2 .$$

Note that the GREG estimate above is *not* the same in general as the estimate obtained by substituting $d_j y_j$ for y_j in a “standard” *GREG* estimate for a population total. This simple “substitution” estimate is model-biased, as shown in 3.1.7 above.

3.2 Estimation of change

Most business surveys are continuing surveys. That is, the survey is repeated monthly, quarterly, annually or with some other fixed frequency. An important reason for doing this is to estimate the change in population quantities from one survey period to the next. This estimation would be relatively straightforward if the target population and the survey sample remained the same from one period to the next. Unfortunately, this is almost never the case. Methods for coping with the complications caused by sample and population change over time are discussed below.

To keep notational complexity to a minimum we restrict ourselves to change in a finite population total between two time points. Let t_1 denote the population total of a survey variable Y at time T_1 and let t_2 denote the corresponding total at time T_2. The values of Y at time T_1 will be denoted y_{1j} and the values of Y at time T_2 will be denoted y_{2j}. The aim is to estimate either the absolute change $\delta = t_2 - t_1$ or the relative change $\phi = (t_2 - t_1)/t_1 = t_2/t_1 - 1$.

Real populations are rarely static. Thus, the units making up the population contributing to t_1 will be different from those making up the population contributing to t_2. We put N_u, u = 1, 2 equal to the number of units in the population at time T_u. In many cases there will be considerable overlap between the populations at the two time points. We put C (“continuing”) equal to the set of population units common to both time points. The set of population units contributing to t_1 and not t_2 will be denoted D ("deaths") while the set contributing to t_2 and not t_1 will be denoted B ("births"). Let N_C, N_D and N_B denote the numbers of units in these sets respectively. Then $N_1 = N_C + N_D$ and $N_2 = N_C + N_B$. The "total" population will be denoted as the set of units contributing to either t_1 or t_2 or both. Clearly this contains $N_C + N_\mathrm{D} + N_B$ units.

A similar decomposition of the sample s at times T_1 and T_2 can be defined. Thus s_1 is the sample at time T_1, s_2 is the sample at time T_2, s_c is the sample common to both times, s_d is the set of sample units unique to time T_1 and s_b denotes the sample units unique to time T_2. Note that units in s_c must, by definition, be in C, but units in s_d do not have to be in D, and similarly units in s_b do not have to be in B. We put s_{dD} equal to those units in s_d *and D*, with $s_{dC} = s_d - s_{dD}$. Similarly, we put s_{bB} equal to those units in s_b *and B*, with $s_{bC} = s_b - s_{bB}$.

3.2.1 Linear estimation

Suppose some form of weighted linear estimate of the population total of Y is computed at each time period. These are estimates of the form (u = 1, 2)

$$\hat{t}_u = \sum_{j \in s_u} w_{uj} y_{uj}$$

where the "*L*" and "*s*" subscripts have been dropped for the sake of clarity. The weights w_{uj} are assumed to be calibrated with respect to known population characteristics at time T_u.

An obvious estimate of δ is then the difference $\hat{\delta} = \hat{t}_2 - \hat{t}_1$. Provided the "level" estimate $\hat{t}_u$ is unbiased for t_u, it is clear that $\hat{\delta}$ will also be unbiased for δ. A corresponding estimate for ϕ is then $\hat{\phi} = \hat{t}_2 / \hat{t}_1 - 1$.

Development of design variances for these estimates is complicated by the need to evaluate the design covariance between $\hat{t}_1$ and $\hat{t}_2$. To illustrate, suppose both $\hat{t}_1$ and $\hat{t}_2$ are *HTE*s, and let the indicator I_{uj} denoting sample inclusion/exclusion at time T_u, so the probability of inclusion in sample of population unit j at time T_u is equal to π_{uj}. Let U_1 denote the population at T_1 and U_2 denote the population at T_2. Then

$$\begin{aligned} \mathrm{V}_p\left(\hat{\delta} - \delta\right) &= \mathrm{V}_p\left(\sum_{j \in U_2} \frac{I_{2j} y_{2j}}{\pi_{2j}} - \sum_{j \in U_1} \frac{I_{1j} y_{1j}}{\pi_{1j}} \right) \\ &= \mathrm{V}_p\left(\sum_{j \in B} \frac{I_{2j} y_{2j}}{\pi_{2j}} + \left\{ \sum_{j \in C} \frac{I_{2j} y_{2j}}{\pi_{2j}} - \frac{I_{1j} y_{1j}}{\pi_{1j}} \right\} - \sum_{j \in D} \frac{I_{1j} y_{1j}}{\pi_{1j}} \right) \end{aligned}$$

which can only be expanded further provided the joint distribution of I_{1j} and I_{2j} can be specified for all pairs of units in the "total" population. This is trivial if independent samples are selected at each time period. However, it is far more common that some form of controlled sample rotation scheme is used. In such cases calculation of this variance can be rather complex. For example, Nordberg (1998) sets out the theory for estimation of the design variance of both $\hat{\delta}$ and $\hat{\phi}$ under the SAMU sample co-ordination system used at Statistics

Sweden for the particular case where simple random sampling within strata is employed at each time period. This approach conditions on the realised sample sizes defined by the random sets s_d, s_b, s_c, s_{dD} and s_{bB}, and so is essentially equivalent to the model-based approach outlined below.

A model-based approach to measuring the variance of $\hat{\delta}$ is reasonably straightforward to develop, though notationally cumbersome. We have

$$
\begin{aligned}
\mathrm{V}_\xi\left(\hat{\delta}-\delta\right) = \mathrm{V}_\xi\Bigg(& \sum_{j\in s_{bB}} w_{2j}y_{2j} + \sum_{j\in s_{bC}} w_{2j}y_{2j} + \sum_{j\in s_c}\left(w_{2j}y_{2j} - w_{1j}y_{1j}\right) - \sum_{j\in s_{dC}} w_{1j}y_{1j} \\
& - \sum_{j\in s_{dD}} w_{1j}y_{1j} - \sum_{j\in B} y_{2j} - \sum_{j\in C}\left(y_{2j}-y_{1j}\right) + \sum_{j\in D} y_{1j}\Bigg) \\
= \quad & \sum_{j\in s_{bB}}\left(w_{2j}-1\right)^2\sigma_{2j}^2 + \sum_{j\in s_{bC}}\left\{\left(w_{2j}-1\right)^2\sigma_{2j}^2 + 2\left(w_{2j}-1\right)\rho_j\sigma_{2j}\sigma_{1j} + \sigma_{1j}^2\right\} \\
& + \sum_{j\in s_c}\left\{\left(w_{2j}-1\right)^2\sigma_{2j}^2 + 2\left(w_{2j}-1\right)\left(w_{1j}-1\right)\rho_j\sigma_{2j}\sigma_{1j} + \left(w_{1j}-1\right)^2\sigma_{1j}^2\right\} \\
& + \sum_{j\in s_{dD}}\left(w_{1j}-1\right)^2\sigma_{1j}^2 + \sum_{j\in s_{dC}}\left\{\left(w_{1j}-1\right)^2\sigma_{1j}^2 + 2\left(w_{1j}-1\right)\rho_j\sigma_{2j}\sigma_{1j} + \sigma_{2j}^2\right\} \\
& + \sum_{B\backslash s_{bB}}\sigma_{2j}^2 + \sum_{C\backslash\left(s_c+s_{bC}+s_{dC}\right)}\left\{\sigma_{2j}^2 - 2\rho_j\sigma_{2j}\sigma_{1j} + \sigma_{1j}^2\right\} + \sum_{D\backslash s_{dD}}\sigma_{1j}^2
\end{aligned}
$$

where σ_{uj} denotes the model standard deviation of y_{uj}, and ρ_j denotes the model correlation between y_{1j} and y_{2j}. Note that $B\backslash s_{bB}$ denotes all elements of B that are not in s_{bB}, and so on. Provided units belonging to the various sample components in the above variance are identifiable, we can estimate the model-variance of $\hat{\delta}$ using "plug in" estimates for the various model parameters in this expression.

A "heteroskedasticity" robust estimate of the model-variance of $\hat{\delta}$ can be written down using the theory set out in 2.3.2.10. Define μ_{uj} as the model expectation of y_{uj}, with unbiased estimate $\hat{\mu}_{uj}$. Suppose further that for some known constant h_{uj} we have $\mathrm{E}_\xi\left[h_{uj}\left(y_{uj}-\hat{\mu}_{uj}\right)^2\right] = \mathrm{E}_\xi\left(y_{uj}-\mu_{uj}\right)^2$. Then we can estimate the model-variance of $\hat{\delta}$ by

$$
\begin{aligned}
& \hat{\mathrm{V}}_\xi^D\left(\hat{\delta}-\delta\right) = \\
& \quad \sum_{j\in s_{bB}}\left(w_{2j}-1\right)^2 h_{2j}\left(y_{2j}-\hat{\mu}_{2j}\right)^2 + \sum_{j\in s_{bC}}\left\{\left(w_{2j}-1\right)\sqrt{h_{2j}}\left(y_{2j}-\hat{\mu}_{2j}\right) + \sqrt{h_{1j}}\left(y_{1j}-\hat{\mu}_{1j}\right)\right\}^2 \\
& \quad + \sum_{j\in s_c}\left\{\left(w_{2j}-1\right)\sqrt{h_{2j}}\left(y_{2j}-\hat{\mu}_{2j}\right) + \left(w_{1j}-1\right)\sqrt{h_{1j}}\left(y_{1j}-\hat{\mu}_{1j}\right)\right\}^2 \\
& \quad + \sum_{j\in s_{dC}}\left\{\left(w_{1j}-1\right)\sqrt{h_{1j}}\left(y_{1j}-\hat{\mu}_{1j}\right) + \sqrt{h_{2j}}\left(y_{2j}-\hat{\mu}_{2j}\right)\right\}^2 + \sum_{j\in s_{dD}}\left(w_{1j}-1\right)^2 h_{1j}\left(y_{1j}-\hat{\mu}_{1j}\right)^2 \\
& \quad + \sum_{B\backslash s_{bB}}\hat{\sigma}_{2j}^2 + \sum_{C\backslash\left(s_c+s_{bC}+s_{dC}\right)}\left\{\hat{\sigma}_{2j}^2 - 2\hat{\chi}_{12j} + \hat{\sigma}_{1j}^2\right\} + \sum_{D\backslash s_{dD}}\hat{\sigma}_{1j}^2
\end{aligned}
$$

where $\hat{\sigma}_{uj}^2$, u = 1, 2 and $\hat{\chi}_{12j}$ are model-based estimates of $V_\xi(y_{uj})$ and $C_\xi(y_{1j}, y_{2j})$ respectively. Thus, for the situation considered in 2.3.2.10, we have

$$\begin{aligned}\hat{V}_\xi^D\left(\hat{\delta}-\delta\right) = & \sum_{j\in s_{bB}}\left(w_{2j}-1\right)^2 h_{2j}\left(y_{2j}-\hat{\mu}_{2w}\right)^2 + \sum_{j\in s_{bC}}\left\{\left(w_{2j}-1\right)\sqrt{h_{2j}}\left(y_{2j}-\hat{\mu}_{2w}\right)+\sqrt{h_{1j}}\left(y_{1j}-\hat{\mu}_{1w}\right)\right\}^2 \\ & + \sum_{j\in s_c}\left\{\left(w_{2j}-1\right)\sqrt{h_{2j}}\left(y_{2j}-\hat{\mu}_{2w}\right)+\left(w_{1j}-1\right)\sqrt{h_{1j}}\left(y_{1j}-\hat{\mu}_{1w}\right)\right\}^2 \\ & + \sum_{j\in s_{dC}}\left\{\left(w_{1j}-1\right)\sqrt{h_{1j}}\left(y_{1j}-\hat{\mu}_{1w}\right)+\sqrt{h_{2j}}\left(y_{2j}-\hat{\mu}_{2w}\right)\right\}^2 + \sum_{j\in s_{dD}}\left(w_{1j}-1\right)^2 h_{1j}\left(y_{1j}-\hat{\mu}_{1w}\right)^2 \\ & + \left(N_B-n_{bB}\right)\hat{\sigma}_{2w}^2 + \left(N_D-n_{dD}\right)\hat{\sigma}_{1w}^2 + \left(N_C-n_c-n_{bC}-n_{dC}\right)\left(\hat{\sigma}_{2w}^2-2\hat{\chi}_{12w}+\hat{\sigma}_{1w}^2\right)\end{aligned}$$

where

$$\begin{aligned} h_{uj} &= 1-2\frac{w_{uj}}{N_u}+\frac{1}{N_u^2}\sum_{j\in s_u}w_{uj}^2 \\ \hat{\mu}_{uw} &= N_u^{-1}\sum_{j\in s_u}w_{uj}y_{uj} \\ \hat{\sigma}_{uw}^2 &= \frac{1}{n_u}\sum_{j\in s_u}\frac{\left(y_{uj}-\hat{\mu}_{uw}\right)^2}{h_{uj}}\end{aligned}$$

and

$$\begin{aligned} \hat{\chi}_{12w}^2 &= \frac{1}{n_c}\sum_{j\in s_c}\frac{\left(y_{2j}-\hat{\mu}_{2cw}\right)\left(y_{1j}-\hat{\mu}_{1cw}\right)}{\sqrt{h_{2cj}h_{1cj}}} \\ \hat{\mu}_{ucw} &= \left(\sum_{j\in s_c}w_{uj}\right)^{-1}\sum_{j\in s_c}w_{uj}y_{uj} \\ h_{ucj} &= 1-2\frac{w_{uj}}{\sum_{k\in s_c}w_{uk}}+\frac{\sum_{k\in s_c}w_{uk}^2}{\left(\sum_{k\in s_c}w_{uk}\right)^2}\end{aligned}$$

Turning now to $\hat{\phi}$, we note that a first order approximation to its model-variance can be written down using a Taylor series argument. This is

$$V_\xi\left(\hat{\phi}-\phi\right)\approx\frac{1}{t_1^2}\left[V_\xi\left(\hat{t}_2-t_2\right)-2\frac{t_2}{t_1}C_\xi\left(\hat{t}_2-t_2,\hat{t}_1-t_1\right)+\left(\frac{t_2}{t_1}\right)^2 V_\xi\left(\hat{t}_1-t_1\right)\right]$$

where

$$V_\xi\left(\hat{t}_1-t_1\right)=\sum_{j\in s_c}\left(w_{1j}-1\right)^2\sigma_{1j}^2+\sum_{j\in s_{dD}}\left(w_{1j}-1\right)^2\sigma_{1j}^2+\sum_{j\in s_{dC}}\left(w_{1j}-1\right)^2\sigma_{1j}^2+\sum_{C\backslash\left(s_c+s_{bC}+s_{dD}\right)}\sigma_{1j}^2+\sum_{D\backslash s_{dD}}\sigma_{1j}^2,$$

$$V_\xi\left(\hat{t}_2-t_2\right)=\sum_{j\in s_{bB}}\left(w_{2j}-1\right)^2\sigma_{2j}^2+\sum_{j\in s_{bC}}\left(w_{2j}-1\right)^2\sigma_{2j}^2+\sum_{j\in s_c}\left(w_{2j}-1\right)^2\sigma_{2j}^2+\sum_{B\backslash s_{bB}}\sigma_{2j}^2+\sum_{C\backslash\left(s_c+s_{bC}\right)}\sigma_{2j}^2 \text{ and}$$

$$\mathrm{C}_{\xi}\left(\hat{t}_2 - t_2, \hat{t}_1 - t_1\right) = \sum_{j \in s_{bC}} \left(w_{2j} - 1\right)\rho_j \sigma_{2j}\sigma_{1j} + \sum_{j \in s_{dC}} \left(w_{1j} - 1\right)\rho_j \sigma_{2j}\sigma_{1j}$$
$$+ \sum_{j \in s_c} \left(w_{2j} - 1\right)\left(w_{1j} - 1\right)\rho_j \sigma_{2j}\sigma_{1j} + \sum_{C \backslash (s_c + s_{bC} + s_{dC})} \rho_j \sigma_{2j}\sigma_{1j}$$

We can estimate the components of this variance using the “heteroskedasticity” robust variance estimation theory set out in 2.3.2.10. Details are omitted, but follow the corresponding development for $\hat{\delta}$ closely.

3.2.2 Estimates of change for functions of population totals

The Taylor series linearisation methods described in 2.4.1 can also be used to estimate the variance of the estimate of change in a function of the population totals at each time point. To illustrate, consider the case where we are interested in the change in the ratio of the population totals of two variables, say Y_a and Y_b. This change is defined as

$$\delta_R = R_2 - R_1 = \frac{\sum_{j \in U_2} y_{a2j}}{\sum_{j \in U_2} y_{b2j}} - \frac{\sum_{j \in U_1} y_{a1j}}{\sum_{j \in U_1} y_{b1j}}$$

Suppose further that these totals are estimated via unbiased linear weighted estimates at each time point. A consistent estimate of δ_R is then

$$\hat{\delta}_R = \hat{R}_2 - \hat{R}_1 = \frac{\sum_{j \in s_2} w_{2j} y_{a2j}}{\sum_{j \in s_2} w_{2j} y_{b2j}} - \frac{\sum_{j \in s_1} w_{1j} y_{a1j}}{\sum_{j \in s_1} w_{1j} y_{b1j}}.$$

The approach described in 2.4.1.1 can be used to “linearise” the estimates of the ratio at each time point. Thus

$$\hat{R}_u \approx \sum_{j \in s_u} w_{uj} z_{uj}$$

where

$$z_{uj} = \frac{y_{auj} - \tilde{R}_u y_{buj}}{\sum_{k \in U_u} \mu_{buk}}$$

$$\tilde{R}_u = \frac{\sum_{j \in U_u} \mu_{auj}}{\sum_{j \in U_u} \mu_{buj}}$$

and μ_{auj}, μ_{buj} represent the expected values of y_{auj} and y_{buj} respectively. Consequently

$$\hat{\delta}_R \approx \sum_{j \in s_2} w_{2j} z_{2j} - \sum_{j \in s_1} w_1 z_{1j}$$

and we can apply the results in 3.2.1 above to estimate the variance of $\hat{\delta}_R$, replacing y_{kj} in the variance estimate formula there by

$$\hat{z}_{uj} = \frac{y_{auj} - \hat{R}_u y_{buj}}{\sum_{j \in s_u} w_{uj} y_{buj}} .$$

Alternatively, either the bootstrap or jackknife approaches to variance estimation can be used. In either case, the "sample" underlying the procedure is the union $s = s_d + s_c + s_b$ of the samples s_1 and s_2. Thus the "drop out 1" jackknife in this case proceeds by deleting one unit at a time from s. See Canty & Davison (1997) for an application of the bootstrapping idea to estimation of change.

3.2.3 Estimates of change in domain quantities

Given linear weighted estimation at each time period, an estimate of the change in domain totals between t_1 and t_2,

$$\delta_d = \sum_{j \in U_2} d_{2j} y_{2j} - \sum_{j \in U_1} d_{1j} y_{1j}$$

where d_{uj} denotes the value of the domain indicator at time T_u for unit j, is $\hat{\delta}_d = \sum_{j \in s_2} w_{2j} d_{2j} y_{2j} - \sum_{j \in s_1} w_{1j} d_{1j} y_{1j}$. As noted in 3.1.7, the level estimate components of $\hat{\delta}_d$ may be biased, and so this estimate of change may be biased as well. To illustrate, consider the stratum model of 3.1.7 with auxiliary variable X defined by stratum indicators plus a size variable Z, and with calibrated weights. Assume further that the coefficient for Z in the regression of Y on X is the same at both time periods. The bias in $\hat{\delta}_d$ is then

$$\mathrm{E}_{\xi}\left(\hat{\delta}_d - \delta_d\right) = \sum_h \theta_h \left(\sum_{j \in s_{ch}} \left(w_{2j} - w_{1j}\right) z_j^{\mathrm{T}} + \left\{ \sum_{j \in s_{bh}} w_{2j} z_j^{\mathrm{T}} - \sum_{j \in s_{dh}} w_{1j} z_j^{\mathrm{T}} \right\} - \left\{ \sum_{j \in B_h} z_j^{\mathrm{T}} - \sum_{j \in D_h} z_j^{\mathrm{T}} \right\} \right) \beta_z$$

where a subscript of h denotes restriction to stratum h. This bias vanishes if θ_h is the same for all h (the condition for the domain estimates at each time period to be unbiased). In general, however, there is little we can say about this bias. One exception is where the births and deaths within a stratum have approximately the same distribution for Z, in which case the third term in braces above should be small. Similarly, if the weights for the common sample within a stratum remain approximately the same from T_1 to T_2, and the incoming sample at time T_2 is chosen so that it "represents" the same proportion of the stratum total of Z as the outgoing sample from T_1, then the first and second terms in braces will also be close to zero and so the bias in $\hat{\delta}_d$ will be small. Variance estimation for a linear weighted $\hat{\delta}_d$ is straightforward. We replace y_{kj} in the variance formulae in the preceding sections by $d_{kj}y_{kj}$. Note that a corresponding modification to the estimate of the expected value μ_{kj} of this variable is also required when computing residuals for use in the variance estimate. Furthermore, since the domain inclusion variable D and the survey variable Y are uncorrelated under ξ (that is, given the values of the auxiliary variable), $\mathrm{V}_\xi\left(d_{kj}y_{kj}\right) = \mathrm{E}_\xi\left[d_{kj}\left(y_{kj} - \mu_{kj}\right)^2\right]$. Applying the model-robust variance estimate developed in 3.2.1 then leads to

$$\begin{aligned}
\hat{\mathrm{V}}_\xi^D\left(\hat{\delta}_d - \delta_d\right) = & \sum_{j \in s_{bB}} \left(w_{2j} - 1\right)^2 h_{2j} d_{2j} \left(y_{2j} - \hat{\mu}_{2j}\right)^2 \\
& + \sum_{j \in s_{bC}} \left\{\left(w_{2j} - 1\right)\sqrt{h_{2j}}\, d_{2j}\left(y_{2j} - \hat{\mu}_{2j}\right) + \sqrt{h_{1j}}\, d_{1j}\left(y_{1j} - \hat{\mu}_{1j}\right)\right\}^2 \\
& + \sum_{j \in s_c} \left\{\left(w_{2j} - 1\right)\sqrt{h_{2j}}\, d_{2j}\left(y_{2j} - \hat{\mu}_{2j}\right) + \left(w_{1j} - 1\right)\sqrt{h_{1j}}\, d_{1j}\left(y_{1j} - \hat{\mu}_{1j}\right)\right\}^2 \\
& + \sum_{j \in s_{dC}} \left\{\left(w_{1j} - 1\right)\sqrt{h_{1j}}\, d_{1j}\left(y_{1j} - \hat{\mu}_{1j}\right) + \sqrt{h_{2j}}\, d_{2j}\left(y_{2j} - \hat{\mu}_{2j}\right)\right\}^2 \\
& + \sum_{j \in s_{dD}} \left(w_{1j} - 1\right)^2 h_{1j} d_{1j} \left(y_{1j} - \hat{\mu}_{1j}\right)^2 + \sum_{B \backslash s_{bB}} \hat{v}_{2j}^2 + \sum_{C \backslash \left(s_c + s_{bC} + s_{dC}\right)} \left\{\hat{v}_{2j}^2 - 2\hat{v}_{12j} + \hat{v}_{1j}^2\right\} + \sum_{D \backslash s_{dD}} \hat{v}_{1j}^2
\end{aligned}$$

where the estimated variances and covariances contributing to the second order (unweighted) terms in this variance estimate are given by $v_{kj}^2 = \hat{\sigma}_{kj}^2 \hat{\theta}_{kj} + \hat{\mu}_{kj}^2 \hat{\theta}_{kj}\left(1 - \hat{\theta}_{kj}\right)$ and $v_{12j} = \hat{\sigma}_{12j}\hat{\theta}_{12j} + \hat{\mu}_{1j}\hat{\mu}_{2j}\left[\hat{\theta}_{12j} - \hat{\theta}_{1j}\hat{\theta}_{2j}\right]$. Here σ_{12j} is the covariance between y_{1j} and y_{2j}, and θ_{12j} is the probability of domain inclusion at both T_1 and T_2. Both these quantities need to be modelled using data from the common sample s_c.

3.3 Outlier robust estimation

Outliers are a common problem in sample surveys, and particularly in business surveys. Given a model ξ for a survey variable *Y*, an outlier is a value for this variable, which is essentially "impossible" under ξ. An outlier is therefore an indication of a breakdown in the model specification for *Y*. Outliers can be both sample and non-sample values. In the latter case, however, they are not observed, and so this misspecification is never identified. In what follows therefore we confine attention to sample outliers. We also assume such outliers are "representative", in the sense that they are not caused by errors in data collection or processing. That is these values are "real" – they are just not at all like the rest of the sample values.

By definition, outliers are rare. Consequently, although their presence in the sample tells us that ξ is misspecified, there are so few of them that there is not enough information to modify the definition of ξ in order to accommodate them. For example, outliers often arise because industry and size characteristics used to define strata are out of date, and so a stratum ends up containing units whose "current" characteristics (and resulting economic performance) are quite unrelated to that of the majority of units in the stratum. If there is a substantial proportion of these incorrectly classified units, then stratum level estimates can be replaced by domain estimates. However, typically there are only a few such outliers, and domain estimation procedures based on these are highly unstable.

There are three basic approaches to dealing with sample outliers. The first is the most common in practice and the least defensible. This is to delete the outliers from the sample. This cannot be justified unless there is strong evidence that the sample outliers are "unrepresentative", being due to incorrect data collection methods or errors introduced in sample processing. The second is to keep the outliers in sample, but to give them weights equal to one. This corresponds to the assumption that the outliers are unique, and that there are no remaining outliers in the nonsampled part of the population. This assumption stabilises the overall sample estimate, but at the cost of a potentially large bias. The theoretically most acceptable option is to keep the outliers in the sample, but to modify them so that their impact on the sample estimate is kept small. In effect, the "normal" sample weight that would be associated with the outlier is kept, but the outlier value is modified to something less extreme.

In the following two sub-sections we discuss approaches to this "value modification". By definition these are model-based. Strictly speaking, outliers are irrelevant from the design-based point of view since this theory makes no assumptions about whether a realised sample value is consistent with an assumed superpopulation model for the population data.

We also restrict ourselves to what are sometimes referred to as "*Y*-outliers", that is where the problem is in the realised *Y*-values of certain sample units. Another class of outliers occur where the *X* values of a few sample units are very distant from the *X*-values of the other sample units. These are "*X*-outliers", and they can have a substantial impact on the stability of the overall sample estimate because of their so-called "leverage". This is typically manifested in outlying sample weights, rather than outlying sample values. There are methods for dealing with such outlying weights (see Chambers, 1997), but since they primarily relate to efficient weighting methods rather than to bias and variance issues under probability sampling, they are not discussed further in this report.

3.3.1 Outlier robust model-based estimation

Robust model-based methods for survey estimation are reviewed in Chambers & Kokic (1993). See also Lee (1995). We assume that the "non-outlier" sample values follow the superpopulation model ξ specified in 2.3.2.1, that is where

$$\begin{aligned} \mathrm{E}_{\xi}(y_j) &= \mu(x_j;\omega) \\ \mathrm{V}_{\xi}(y_j) &= \sigma^2(x_j;\omega) \\ \mathrm{C}_{\xi}(y_j, y_k) &= 0 \quad \text{for } j \neq k \end{aligned}$$

However, the sample data contain a few values that are inconsistent with this model. If we ignore these inconsistencies (that is, include the data as normal), our estimate of the population total of *Y* is typically of the form

$$\hat{t}_{\xi} = \sum_{j\in s} y_j + \sum_{j\notin s} \mu(x_j;\hat{\omega})$$

where $\hat{\omega}$ is an estimate of ω based on the sample data. Typically, in the interests of efficiency, this estimate is based on the application of nonrobust estimation methods like least squares or maximum likelihood. The presence of sample outliers can seriously destabilise this estimate however.

Outliers in the population can be modelled by assuming that the population is in fact a mixture of outliers and non-outliers. That is, the "true" superpopulation model for Y is made up of values drawn from ξ and values drawn from an "outlier" model η. This can be represented as

$$y_j = \delta_j\left(\mu(x_j;\omega) + \varepsilon_{\xi j}\right) + (1-\delta_j)\left(\nu(x_j;\gamma) + \varepsilon_{\eta j}\right)$$

where δ_j is an indicator random variable which determines whether a value is an outlier ($\delta_j = 0$) or a non-outlier ($\delta_j = 1$), and $\varepsilon_{\xi j}$ and $\varepsilon_{\eta j}$ are zero mean random variables such that $\mathrm{V}_\xi(\varepsilon_{\xi j}) = \sigma^2(x_j;\omega)$ and $\mathrm{V}_\eta(\varepsilon_{\eta j}) = \tau^2(x_j;\gamma)$, with $\tau^2(x_j;\gamma) >> \sigma^2(x_j;\omega)$. If we further assume that the random variables δ_j and $\varepsilon_{\xi j}$, $\varepsilon_{\eta j}$ are independent of one another, then the "true" population model is such that

$$\mathrm{E}(y_j) = \mu(x_j;\omega) + (1-\pi_j)(\nu(x_j;\gamma) - \mu(x_j;\omega))$$

and

$$\mathrm{V}(y_j) = \pi_j\sigma^2(x_j;\omega) + (1-\pi_j)\tau^2(x_j;\gamma) + \pi_j(1-\pi_j)(\mu(x_j;\omega) - \nu(x_j;\gamma))^2$$

where $\pi_j = \Pr(\delta_j = 1)$. The bias in $\hat{t}_\xi$ is therefore

$$\mathrm{E}(\hat{t}_\xi - t) = \sum_{j\notin s}\mathrm{E}\{\mu(x_j;\hat{\omega}) - \mu(x_j;\omega)\} - \sum_{j\notin s}(1-\pi_j)\{\nu(x_j;\gamma) - \mu(x_j;\omega)\}.$$

The first term on the right hand side above will be essentially zero provided the method for calculating $\hat{\omega}$ can be made outlier robust (for example if the sample outliers have little or no influence on its value). This leads to the estimate

$$\hat{t}_\xi^* = \sum_{j\in s} y_j + \sum_{j\notin s}\mu(x_j;\hat{\omega}_{robust})$$

where $\hat{\omega}_{robust}$ is the outlier robust estimate of ω (this may be simply the estimate of ω obtained after outliers are deleted from the sample). In any case we shall assume that

$$\mathrm{E}\{\mu(x_j;\hat{\omega}_{robust}) - \mu(x_j;\omega)\} \approx 0$$

so the bias of $\hat{t}_\xi^*$ becomes

$$\mathrm{E}(\hat{t}_\xi^* - t) \approx -\sum_{j\notin s}(1-\pi_j)\{\nu(x_j;\gamma) - \mu(x_j;\omega)\}.$$

This bias can still be substantial. Consequently, it is generally insufficient to replace nonrobust parameter estimates by robust parameter estimates when dealing with outliers in sample survey data. However, since

$$\mathrm{E}\left(\sum_{j\notin s}\{y_j-\mu(x_j;\omega)\}\right)=\sum_{j\notin s}(1-\pi_j)\{\nu(x_j;\gamma)-\mu(x_j;\omega)\},$$

one can see that this bias can be estimated by estimating the nonsample total of the residuals generated by ξ. It follows $\hat{t}_\xi^*$ can be corrected by subtracting this estimated bias.

One problem with this estimated bias correction is that the presence of sample outliers can make it very unstable. Chambers (1986) therefore suggested that this correction be "robustified" as well, leading to the modified estimate

$$\hat{t}_\xi^{**}=\sum_{j\in s}y_j+\sum_{j\notin s}\mu(x_j;\hat{\omega}_{robust})+\sum_{j\in s}m_j\sigma(x_j;\hat{\omega}_{robust})\psi\left(\frac{y_j-\mu(x_j;\hat{\omega}_{robust})}{\sigma(x_j;\hat{\omega}_{robust})}\right)$$

where m_j is a suitable chosen weight of order $O((N-n)/n)$ and ψ is a bounded skew-symmetric function which determines the "influence" of the sample residuals on the bias correction.

In the case where $\hat{t}_\xi$ is a general linear weighted estimate, defined by sample weights $\{w_{js}\}$, $\hat{t}_\xi^{**}$ is given by

$$\hat{t}_w^{**}=\sum_{j\in s}y_j+\sum_{j\notin s}\mu(x_j;\hat{\omega}_{robust})+\sum_{j\in s}(w_{js}-1)\sigma(x_j;\hat{\omega}_{robust})\psi\left(\frac{y_j-\mu(x_j;\hat{\omega}_{robust})}{\sigma(x_j;\hat{\omega}_{robust})}\right).$$

A GREG version of $\hat{t}_\xi^{**}$ can also be written down. This is

$$\hat{t}_\pi^{**}=\sum_{j\in U}\mu(x_j;\hat{\omega}_{\pi robust})+\sum_{j\in s}\frac{\sigma(x_j;\hat{\omega}_{\pi robust})}{\pi_j}\psi\left(\frac{y_j-\mu(x_j;\hat{\omega}_{\pi\, robust})}{\sigma(x_j;\hat{\omega}_{\pi\, robust})}\right)$$

where $\hat{\omega}_{\pi robust}$ denotes a design consistent estimate of a *FPP* which is itself a robust estimate of ω.

Choice of the influence function ψ is typically left to the user. A wide variety of such functions are available in the statistics literature (Huber, 1981). In general a "safe" choice seems to be the Huber influence function $\psi(t)$ = sgn(t) × min(abs(t), c), with c not too small, say c = 6. This allows the sample outliers to have some say in the bias correction term, but not enough to destabilise it completely.

In general, none of the above versions of the robust estimate $\hat{t}_\xi^{**}$ is unbiased. However, their mean squared error properties are typically superior to both $\hat{t}_\xi$ and the naive robust estimate $\hat{t}_\xi^*$.

Variance estimation for $\hat{t}_\xi^{**}$ is complicated by this bias property, as well as by the intrinsic nonlinearity of the estimate. Chambers & Dorfman (1994) report on the use of the bootstrap to estimate confidence intervals for robust estimates like $\hat{t}_\xi^{**}$. In general, they found that the bootstrap variance estimates could not handle the bias, leading to actual confidence interval coverage that was less than nominal coverage.

The estimate $\hat{t}_{\xi}^{**}$ is motivated by what is sometimes referred to as a “gross error model” for the population outliers. This model is questionable when the outliers are the consequence of a long tailed error distribution for *Y* rather than contamination. Here the outliers arise because of misspecification of ξ. Chambers *et al.* (1993) suggested that in this case one should add a nonparametric bias correction term to $\hat{t}_{\xi}$. Under long-tailed alternatives to ξ, it is wise to “robustify” this nonparametric correction term so that, like the parametric correction term used in $\hat{t}_{\xi}^{**}$, it is relatively unaffected by a few very extreme sample values. This leads to the estimate

$$\hat{t}_{\xi}^{+} = \hat{t}_{\xi} + \sum_{j \notin s} \hat{B}_j$$

where $\hat{B}_j$ is the fitted value at x_j of a robust nonparametric smooth of the sample residuals $r_k = y_k - \mu(x_k; \hat{\omega})$, $k \in s$. In the empirical study reported in Chambers & Dorfman (1994), this estimate, based on a Huber-type local linear smoother, performed extremely well, recording both a low bias and a low mean squared error. Bootstrap confidence intervals based on $\hat{t}_{\xi}^{+}$ also had the best coverage properties of all the robust estimates considered in that reference.

3.3.2 Winsorisation-based estimation

As has been noted a number of times before, the use of sample weighted estimates is common in business surveys. Consequently, there is a demand for robust estimation methods that can (at least nominally) fit into this linear estimation framework. The model-based robust estimation methods described above are not easily computed in this way. An alternative method that fits naturally into this framework and has good outlier robustness properties is the so-called winsorisation approach. Under this method, outlying sample *Y*-values are modified so they are no longer outlying, and the linear weighted estimate is then calculated using these modified values.

More precisely, since any linear weighted estimate of a population total can be expressed as

$$\hat{t}_L = \sum_{j \in s} w_{js} y_j = \sum_{j \in s} y_j + \sum_{j \in s} (w_{js} - 1) y_j ,$$

winsorisation proceeds by replacing an outlying y_j value in the second term on the right hand side above by a less outlying value. In particular, the winsorised estimate can be written

$$\hat{t}_L^* = \sum_{j \in s} w_{js} y_j = \sum_{j \in s} y_j + \sum_{j \in s} (w_{js} - 1) [y_j I(L_j \le y_j \le U_j) + L_j I(y_j < L_j) + U_j I(y_j > U_j)]$$

where $I(\cdot)$ denotes an indicator function which takes the value 1 if its argument is true and is zero otherwise and L_j, U_j are lower and upper bounds for the *Y*-value of population unit $j \in s$.

Determination of these bounds depends on the superpopulation model ξ for *Y*. As usual we assume the general superpopulation model of 2.3.2.1. That is, we assume the mean and variance of y_j under ξ are given by $\mu(x_j;\omega)$ and $\sigma^2(x_j;\omega)$ respectively, where ω is an unknown parameter. In many business survey applications, *Y* is intrinsically positive, and so L_j is set to zero. This is referred to as one-sided winsorisation. For this case Kokic & Smith (1999a) parameterise the upper bound U_j in terms of a single parameter *U*, via

$$U_j = \mu(x_j;\hat{\omega}) + \frac{U}{w_{js} - 1}$$

where $\mu(x_j;\hat{\omega})$ is an unbiased estimate of the expected value of y_j under ξ. They then develop procedures for choosing *U* in order to minimise the mean squared error of $\hat{t}_L^*$ under ξ. These procedures require access to historical survey information in order to estimate ω. Empirical results quoted in their paper indicate substantial gains from winsorisation in surveys of "outlier prone" populations.

A problem with one-sided winsorisation is that, by construction, the resulting estimate has a negative bias. Typically, estimation is carried out separately in various strata and these estimates are then added to give an overall population estimate. If winsorisation is applied within each stratum (that is *U* above is determined separately for each stratum in order to minimise mean squared error at stratum level), then the overall population estimate may have a substantial negative bias caused by summation of the individual stratum biases. Thus, although the individual stratum level estimates are well behaved, the overall estimate may have an unacceptable level of error. On the other hand, if *U* is determined at population level (that is, the same *U* in all strata), then this may lead to stratum level estimates that are unacceptable.

In a subsequent paper (Kokic & Smith, 1999b) have extended their methodology so that both lower and upper bounds are determined in such a way as to ensure that the winsorised estimate has minimum variance under ξ subject to it being (approximately) unbiased under this model. Their parameterisations for U_j and L_j in this case are

$$U_j = \mu(x_j;\tilde{\omega}) + \sigma(x_j;\hat{\omega}_{robust})U$$

and

$$L_j = \mu(x_j;\tilde{\omega}) - \sigma(x_j;\hat{\omega}_{robust})L$$

where $\tilde{\omega}$ is an *independent* unbiased estimate of ω (for example obtained from historical survey data) and $\sigma(x_j;\hat{\omega}_{robust})$ is an outlier robust estimate of the standard deviation of y_j under ξ. The cut-off parameters *U* and *L* are then chosen in order to minimise the model variance of $\hat{t}_L^*$ subject to it having zero model bias. It turns out that these optimal values depend on solution of two differential equations defined by the common distribution *F* of the standardised residuals $r_j = (y_j - \mu(x_j;\omega)/\sigma(x_j;\omega))$. These are

$$F(-L)dL = (1 - F(U))dU$$

and

$$U + L = (1 + dU/dL)\{f(U) + f(-L)dL/dU\}$$

where f is the density corresponding to F. Empirical results reported in Kokic & Smith (1999b) indicate that this two-sided winsorised estimate overcomes the "cumulative bias" problem described above for one-sided winsorisation, while still retaining the outlier robustness properties associated with the winsorisation idea.

Provided $\tilde{\omega}$ (and hence U_j and L_j) is based on independent historical information, variance estimation for $\hat{t}_L^*$ is straightforward, since the methods described in previous sections of this report can be applied, with y_j replaced by its winsorised value

$$y_j^* = y_j I(L_j \leq y_j \leq U_j) + L_j I(y_j < L_j) + U_j I(y_j > U_j)$$

When historical data are not available, it is unclear how one can proceed to determine L and U above. One possibility is to use *cross-validation*, using part of the sample to determine $\tilde{\omega}$ and the rest to determine L and U, and then repeating this process for a set of nonoverlapping subsamples which essentially cover the original sample. The final values of L and U are then obtained as averages of these subsample-based estimates. The properties of this approach are unknown at the time of writing.

3.4 Variance estimation for indices

Many key official statistics are presented in the form of indices, themselves calculated using estimates derived from a number of sources, both surveys and administrative systems. The purpose of this section is to briefly outline methodology for variance estimation for such statistics. To provide a focus for this discussion, the case of variance estimation for the UK Index of Production (IoP) will be considered. For a more comprehensive assessment, see Kokic (1998).

The IoP is an economic indicator produced by United Kingdom's Office for National Statistics (ONS). It is a monthly index of the total volume of industrial output (or production). It covers the Mining, Manufacturing and Agricultural sectors of the UK economy and is currently based to 1990 prices. It is one of the main indicators of economic growth within the UK.

The IoP is obtained by combining several different sources of data. By far the most significant source is ONS surveys. These include the Monthly Production Inquiry (MPI), Producer Price Index (PPI) and the Quarterly Stocks Inquiry (QSI). Other data used in its construction include the Export Price Deflator (EPD), which is currently derived from a combination of data collected by ONS and by UK Customs and Excise, and additional data on the oil, gas, electricity and mining industries from the UK Department of Trade and Industry, and on food production from the UK Ministry of Agriculture, Fisheries and Food.

The IoP is first constructed within industry groups at the 4-digit standard industrial classification (SIC) level (Central Statistical Office, 1992). Let I_{0Th} be the IoP estimate for time period T relative to a reference period 0 in industry group h. Higher level estimates are produced by taking weighted averages of these IoP estimates, where the weights are determined by the value added in the base year (estimated from the Annual Business Inquiry survey). Thus the overall index I_{0T} is given by

$$I_{0T} = \left(\sum_h I_{0Th} w_{0h} \right) \left(\sum_h w_{0h} \right)^{-1}$$

where w_{0h} is a "value added" weight for industry h. The relative change in the IoP between time periods T_1 and T_2 may be written as

$$I_{T_1T_2} = \sum_h I_{0T_2h} w_{0h} \Big/ \sum_h I_{0t_1h} w_{0h} \,.$$

From now on, except where necessary for clarity, we shall only make reference to one base year, a single reference period T and one 4–digit industry h, and so for simplicity the subscripts 0, T and h will be dropped. The process of index construction can be broken down into a number of distinct steps.

Step 1: *Construction of the combined price deflator*. A price deflator for home (that is, domestic) sales is estimated from PPI data, and another for export sales is estimated from EPD data. The inverses of these deflators estimate the average price increase from the base year for commodities produced and sold by all members of a given industry. The combined deflator is a harmonic mean of these home and export price deflators, weighted by total home sales and total export sales, both estimated from MPI data. It is defined by

$$\hat{D} = \left(\frac{\hat{S}_{\text{home}}}{\hat{D}_{\text{home}}} + \frac{\hat{S}_{\text{export}}}{\hat{D}_{\text{export}}} \right)^{-1} \hat{S}$$

where $\hat{D}_{\text{home}}$ is the estimated home price deflator (from PPI), $\hat{D}_{\text{export}}$ is the estimated export price deflator (from EPD), $\hat{S}_{\text{home}}$ is estimated home sales (from MPI), $\hat{S}_{\text{export}}$ is estimated export sales (from MPI) and $\hat{S} = \hat{S}_{\text{home}} + \hat{S}_{\text{export}}$.

Step 2: *Construction of a deflated weighted sales index*. This index represents the relative increase in real terms of sales in the current month compared to the base year. For this purpose sales are split between merchanted goods and non-merchanted goods. Merchanted goods are those products "sold on" by a business without being subjected to a manufacturing process. The index is defined by

$$I_{\text{sales}} = \left\{ \left(\frac{\hat{S} - \hat{S}_m}{\hat{S}} \right) \left(\frac{\hat{S} - \hat{S}_m}{\hat{G} - \hat{G}_m} \right) + \left(\frac{\hat{S}_m}{\hat{S}} \right) \frac{\hat{S}_m}{\hat{G}_m} \right\} \frac{1}{\hat{D}}$$

where $\hat{S}_m$ is the estimate of sales of merchanted goods (from MPI), $\hat{G}_m$ is the estimate of monthly average sales of these goods in the base year, and $\hat{G}$ is the corresponding estimate of monthly average sales of all goods in the base year.

Step 3: *Creation of a benchmark sales index*. This index is calculated by a linear transformation of the deflated weighted sales index. A multiplicative adjustment is used to ensure that the index meets certain (externally imposed) constraints for publication, and additive tuning constants are used for minor adjustments where the index value does not follow patterns expected in the relevant industry. The index value that is produced is therefore

$$I = I_{\text{sales}} \frac{c}{\hat{d}} + a$$

where c is the constraining factor, $\hat{d}$ is the monthly average of the deflated weighted sales index in the base year and a is the tuning constant.

The final value of the IoP is obtained after carrying out a further additive stock adjustment to the benchmark sales index above. This is then seasonally adjusted before publication, using X11-ARIMA.

Taylor series linearisation and bootstrap methods for estimating the variance of the non-seasonally adjusted IoP are discussed in Kokic (1998). Both are based on the assumption that $\hat{d}$ is approximately one and

$$\frac{\hat{S} - \hat{S}_m}{\hat{S}} \approx \frac{\hat{G} - \hat{G}_m}{\hat{G}}$$

so

$$I_{\text{sales}} \approx \frac{\hat{S}}{\hat{G}\hat{D}} = \frac{1}{\hat{G}} \left(\frac{\hat{S}_{\text{home}}}{\hat{D}_{\text{home}}} + \frac{\hat{S}_{\text{export}}}{\hat{D}_{\text{export}}} \right).$$

It follows that

$$\mathrm{V}(I) \approx c^2\, \mathrm{V}(I_{\text{sales}})$$

where V(I_{sales}) can be estimated via Taylor series linearisation or bootstrapping. In the former case this leads to the estimate

$$\hat{\mathrm{V}}(I_{\text{sales}}) \approx \frac{1}{\hat{G}^2} \left\{ \left(\frac{\hat{S}_{\text{home}}}{\hat{D}_{\text{home}}} + \frac{\hat{S}_{\text{export}}}{\hat{D}_{\text{export}}} \right)^2 \frac{\hat{\mathrm{V}}(\hat{G})}{\hat{G}^2} + \frac{\hat{S}^2_{\text{home}} \hat{\mathrm{V}}(\hat{D}_{\text{home}})}{\hat{D}^4_{\text{home}}} + \frac{\hat{S}^2_{\text{export}} \hat{\mathrm{V}}(\hat{D}_{\text{export}})}{\hat{D}^4_{\text{export}}} \right.$$
$$\left. + \frac{\hat{\mathrm{V}}(\hat{S}_{\text{home}})}{\hat{D}^2_{\text{home}}} + 2\frac{\hat{\mathrm{C}}(\hat{S}_{\text{home}}, \hat{S}_{\text{export}})}{\hat{D}_{\text{home}}\hat{D}_{\text{export}}} + \frac{\hat{\mathrm{V}}(\hat{S}_{\text{export}})}{\hat{D}^2_{\text{export}}} \right\}$$

where a “hat”, as usual, denotes an estimate, and we have used the fact that $\hat{G}$, $\hat{D}_{\text{home}}$, $\hat{D}_{\text{export}}$ and $\left(\hat{S}_{\text{home}}, \hat{S}_{\text{export}}\right)$ are uncorrelated estimates, being based on data collected at two different time points and from three different sources (PPI, EPD and MPI).

A parametric bootstrap estimate of the variance of I_{sales} is also easily computed. This involves sampling with replacement from the large sample approximate joint distribution of $\hat{G}$, $\hat{D}_{\text{home}}$, $\hat{D}_{\text{export}}$ and $\left(\hat{S}_{\text{home}}, \hat{S}_{\text{export}}\right)$. Using a subscript of b to denote such a draw, we have

$$I_b = cI_{\text{sales},b}$$

where

$$I_{\text{sales},b} = \frac{1}{\hat{G}_b}\left(\frac{\hat{S}_{\text{home},b}}{\hat{D}_{\text{home},b}} + \frac{\hat{S}_{\text{export},b}}{\hat{D}_{\text{export},b}}\right)$$

and

$$\begin{aligned}
\hat{D}_{\text{home},b} &\overset{IID}{\sim} N\left(\hat{D}_{\text{home}}, \hat{\mathrm{V}}\left(\hat{D}_{\text{home}}\right)\right) \\
\hat{D}_{\text{export},b} &\overset{IID}{\sim} N\left(\hat{D}_{\text{export}}, \hat{\mathrm{V}}\left(\hat{D}_{\text{export}}\right)\right) \\
\begin{pmatrix} \hat{S}_{\text{home},b} \\ \hat{S}_{\text{export},b} \end{pmatrix} &\overset{IID}{\sim} N\left(\begin{pmatrix} \hat{S}_{\text{home}} \\ \hat{S}_{\text{export}} \end{pmatrix}, \begin{bmatrix} \hat{\mathrm{V}}\left(\hat{S}_{\text{home}}\right) & \hat{\mathrm{C}}\left(\hat{S}_{\text{home}}, \hat{S}_{\text{export}}\right) \\ \hat{\mathrm{C}}\left(\hat{S}_{\text{home}}, \hat{S}_{\text{export}}\right) & \hat{\mathrm{V}}\left(\hat{S}_{\text{export}}\right) \end{bmatrix}\right) \\
\hat{G}_b &\overset{IID}{\sim} N\left(\hat{G}, \hat{\mathrm{V}}\left(\hat{G}\right)\right)
\end{aligned}$$

where $\overset{IID}{\sim}$ denotes a random draw from the indicated distribution. Given B simulated values I_b generated according to this model, the bootstrap variance estimate for I is therefore

$$\hat{\mathrm{V}}_{bootstrap}(I) = \frac{1}{B-1}\sum_{b=1}^{B}\left(I_b - B^{-1}\sum_{b=1}^{B} I_b\right)^2 .$$

In the simulation study reported in Kokic (1998), this approach and Taylor series linearisation led to comparable variance estimates.

3.5 Conclusions

This chapter has extended the theory for estimation and sample error variance estimation introduced in the previous chapter to four important special cases that occur often in business surveys. These are estimation for domains, estimation of change, estimation in the presence of sample outliers and estimation of indices. All four situations require careful application of the theory developed in chapter 2, with an emphasis perhaps on the use of model-based ideas to highlight issues relating to the overall quality of the estimates produced.

High quality domain estimation is a fundamental objective of most business surveys. For example, it is a basic requirement for any survey where the industry and size classification on the frame is out of date. In section 3.1, therefore, we set out the relevant theory for this objective. It is interesting to note that if one treats domain membership on the same basis as any other survey variable, then standard design-based and model-based estimation methods essentially result in the same inference. However, the introduction of extra information about the domains (for example their sizes) can only be easily accommodated from a model-based perspective, though even here there is some debate about exactly how this should be done. Consequently we recommend that when methodology for domain estimation is used in a survey, careful attention is paid to informing the user of these estimates about the method of computation, plus the basis of the sampling variance calculations (that is, whether they are conditional on domain membership in the sample or not).

Estimation of change based on data obtained from typically overlapping samples is another common feature of business surveys. One could in fact claim that such a measure of change is in fact the key objective of most such surveys. In this context we have indicated the manner in which variance estimates both for absolute as well as relative change need to be calculated. Of necessity, these calculations are rather complex involving the integration of survey data from two (and sometimes more) sources. At present we are not aware of any software that can "automatically" carry out these calculations, so the appropriate methodology needs to be "custom programmed" into a survey data analysis system. The theory set out in section 3.2 should be helpful in this regard.

Sample outliers are a perennial problem in business surveys and form the focus of the discussion in section 3.3. Here it suffices to note that a consensus on dealing with these units has yet to be reached, in large part due to the fact that the concept of what constitutes an "outlier" remains the object of debate. The winsorisation methods discussed in section 3.3.2 offer considerable promise and are the subject of current research. Again, use of these methods will generally stabilise the estimated variance of a survey estimate, but at the cost of some increase in bias. This trade-off is typically advantageous if one's main concern is "tracking" the behaviour of the non-volatile part of the target population. In doing so, one should take care, however, to ensure that sample units identified and downweighted as outliers should be investigated and the reasons for their outlying values established. At the end of the day the presence of outliers is a symptom of a badly specified model for the population, and so the information they provide needs to be used to update and improve sample estimation and inference procedures.

Finally, in section 3.4 we tackle the issue of variance estimation for an index calculated on the basis of continuing survey data. Because of the wide variety of such indices in use, we have chosen to confront this problem via discussion of one particular index, the UK Index of Production, and to show how the "complex statistics" methodology discussed in section 2.4 can be adapted to the problem of estimating the sampling variability of this index. The methods (Taylor series linearisation, bootstrapping) we describe are generally applicable to any index, however.

4 Sampling errors under non-probability sampling

David Draper and Russell Bowater[2], University of Bath

4.1 Introduction

In Chapter 2 we examined sampling errors arising from probability sampling. In the random-sampling approach to surveys – and assuming (as we did in Chapter 2) (a) that the target and survey populations coincide, so that one may speak without confusion simply about the population, and (b) that the available frame is perfect – the sampling method is assumed to treat the N population units in such a way that every unit has a non-zero probability of inclusion in the sample.

Continuing the notation in Chapter 2, let y be an outcome variable of interest and define the sample inclusion indicators $I_j = 1$ if unit j is in the sample and 0 otherwise. Probability sampling makes the I_j random variables, so that it is meaningful to speak of the inclusion probability for unit j, $\pi_j = P_p(I_j = 1)$, and the joint inclusion probability for units j and k, $\pi_{jk} = P_p(I_j = 1, I_k = 1)$. Here, as in Chapter 2, the subscript p denotes probability as defined by the (design-based) hypothetical process of repeated random sampling.

As Särndal *et al.* (1992) note, a probability sampling design for which the following two properties hold,

$$\begin{aligned} &\pi_j > 0 \text{ for all } 1 \le j \le N \\ &\pi_{jk} > 0 \text{ for all } 1 \le j \ne k \le N, \end{aligned} \tag{4.1}$$

and for which all of the π_j and π_{jk} are known, is called *measurable*. The first of the conditions in (4.1) (together with the stipulation that the π_j are known) is necessary and sufficient for obtaining a design-unbiased estimator of the population total $t = \sum_{j=1}^{N} y_j$ and the second condition permits the calculation of a (nearly) design-unbiased estimate of the variance of the sample error distribution for estimators of t.

From the design-based point of view, measurable probability sampling designs are thus clearly desirable (Neyman 1934, Cochran 1977), and – as noted in Chapter 2 – probability sampling also provides an important degree of robustness from the model-based perspective. Despite this, non-measurable sampling is frequently employed in some fields even today: *samples of convenience*, in which the π_j are unknown because no attempt was made to choose the sample randomly, are ubiquitous in medicine and the social sciences (Draper 1995b gives several examples of such samples), and probability-sampling designs in which

[2] We are grateful to Ray Chambers (University of Southampton), Eva Elvers (Statistics Sweden) and Paul Smith (UK Office for National Statistics) for comments and references, and to Paul Smith for some suggested text fragments. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people responsible for any errors or omissions that may be present.

some of the π_{jk} are zero (such as stratified random sampling with only one sample unit in one or more strata) can occur in practice.

Non-probability sampling is also sometimes used in business surveys (see Särndal *et al*. 1992 and Lessler & Kalsbeek 1992 for examples). As noted in Eurostat (1996:04), this can occur when there is no readily available sampling frame, or when the survey is voluntary. In this chapter in Sections 4.2–4.5 we consider each of the four leading potential instances of non-probability sampling in business surveys - *voluntary sampling*, *quota sampling*, *judgemental sampling*, and *cut-off sampling*. In Section 4.6 we provide some conclusions, including brief recommendations on best practice and their implications for model quality reports.

It is perhaps worth emphasising at the outset (a) that one of the main problems posed by non-probability sampling is *bias* (as defined in Chapter 2), and (b) that bias is qualitatively different from the kinds of errors that can arise with (small) random samples. In the latter case (design) unbiasedness is guaranteed, in the usual long-run-average sense (see chapters 2 and 3), by the randomisation, and we have only to take larger samples to diminish the likely amount by which our estimates will differ from their true values. Bias is more insidious: it will not go away with increasing sample size, because repeating a biased method of data collection on a larger scale merely perpetuates the bias. Thus there is a major burden on anyone who wishes to use a non-probability sampling method, namely demonstrating that any bias induced by the sampling method can be largely diminished by adjustments such as *poststratification* (to be described in Section 4.2). Even if bias is largely controlled, the unavailability (or non-positivity) of the π_j and/or π_{jk} may create serious problems for accurate uncertainty assessment.

4.2 Voluntary sampling

Voluntary sampling arises when, for example, businesses are requested, but not required, to take part in a survey, and the survey results are based just on the data received from the companies who *choose* to respond. The choice of whether to participate thus makes the sample non-probability-based: even if one wished to acknowledge uncertainty, before the data arrive, about which companies will respond by regarding the sample inclusion indicators I_j as random, the inclusion probabilities π_j are rendered unknown by the choice mechanism. As with quota sampling (Section 4.3), the result can range from highly accurate to highly inaccurate, depending on the (possibly unknown) degree to which the volunteer units represent the population in all relevant respects. Any bias that arises from failure of the voluntary sample to match the population in this way is an example of *selection bias* (see Freedman *et al*. 1998 for a discussion), in which the self-selection mechanism is correlated with the outcome of interest and some or all of its most important predictors.

An example of voluntary sampling is provided by the *Stocks Inquiry* business survey, conducted by the UK Office for National Statistics (ONS). This survey has both a monthly voluntary component and a quarterly component based on probability sampling: random samples of companies are (a) chosen, (b) required to provide quarterly data, and (c) *requested* (in addition) to provide monthly data, so that the companies providing voluntary monthly information form a self-selected subset of the probability sample. In practice about 30% of the sampled companies choose to supply the voluntary data. Note that this type of sample could equally well be described as a probability sample (a) with a voluntary sub-sample or (b) with a high degree of (almost certainly) non-ignorable non-response (see chapter 8 and section 9.7).

Period	Industry 1			Industry 2			Industry 3		
	P	V	$\hat{B}$	P	V	$\hat{B}$	P	V	$\hat{B}$
‘97/Q1	3,420	5,425	+2,005	38,011	38,905	+894	26,617	61,534	+34,917
‘97/Q2	3,456	6,148	+2,692	40,502	43,271	+2,769	27,439	62,990	+35,551
‘97/Q3	3,455	6,008	+2,553	36,940	44,170	+7,320	26,059	59,931	+33,872

Table 4.1 Estimates based on the Probability (P) and voluntary (V) samples, by industry and period, for work-in-progress **Opening** stocks. All figures are in £000. $\hat{B}$ = estimated bias.

Available variables in the analysis we present here include industry group number (four-digit SIC92; we focus here on only 3 industries, coded 1-3); period of return from 01/1997 to 09/1997; register employment and (VAT) turnover (in £000) based on data gathered roughly 3 months previously; and the Opening and Closing stocks (in £000) for each of three categories: materials, work in progress, and finished goods. The numbers of companies involved in the voluntary and probability samples in this period were 77-87 and 261-275, respectively, varying a bit from quarter to quarter. We concentrate here on the work-in-progress stocks (results for the other two categories were similar). For ease of exposition (a) we present results only on the 77 and 226 companies in the voluntary and probability samples with complete data at all time points relevant to the analyses below, and (b) we analyse the data as if the probability sample had been a simple random sample (in fact it was a stratified random sample; the points we wish to make in this section come through more clearly without the extra issue of re-weighting the probability sample back to the population).

Period	Industry 1			Industry 2			Industry 3		
	P	V	$\hat{B}$	P	V	$\hat{B}$	P	V	$\hat{B}$
‘97/Q1	3,456	6,148	+2,692	40,502	43,271	+2,769	27,439	62,990	+35,551
‘97/Q2	3,455	6,008	+2,553	36,940	44,170	+7,320	26,059	59,931	+33,872
‘97/Q3	3,898	7,828	3,930	39,356	49,605	+10,249	24,627	56,638	+32,011

Table 4.2 Estimates based on the Probability (P) and voluntary (V) samples, by industry and period, for work-in-progress **Closing** stocks. All figures are in £000. $\hat{B}$ = estimated bias.

Some indication of the biases that could arise from basing inferences on the voluntary monthly samples is provided by a direct comparison between the monthly and quarterly data in each of the three periods 01-03/97, 04-06/97, and 07-09/97 that were common to both surveys (for comparability between the monthly and quarterly series, the opening and closing of the first quarter of 1997 were taken to be 01/97 and 03/97 for the voluntary series, and analogously for the other quarters). Table 4.1-Table 4.3 present sample estimates by industry and period for work-in-progress Opening, Closing, and (Closing – Opening) stocks in each of these three quarters. Within each industry code, probability (*P*) and voluntary (*V*) estimates are given, and – since we are taking the probability-sampling results to be (design) unbiased – the estimated bias $\hat{B} = V - P$ from the voluntary data may also be calculated. It is evident from these tables (a) that the voluntary results for both opening and closing stocks are enormously biased on the high side, and (b) that much – though by no means all – of this bias cancels in the subtraction when producing the (Closing – Opening) stocks estimates, which are the principal outcomes of interest.

Period	Industry 1			Industry 2			Industry 3		
	P	V	$\hat{B}$	P	V	$\hat{B}$	P	V	$\hat{B}$
‘97/Q1	36	723	+687	2,491	4,366	+1,875	822	1,456	+634
‘97/Q2	-1	-140	-139	-3,562	899	+4,461	-1,380	-3,059	-1,679
‘97/Q3	443	1,820	+1,377	2,416	5,435	+3,019	-1,432	-3,293	-1,861

Table 4.3 Estimates based on the Probability (*P*) and voluntary (*V*) samples, by industry and period, for work-in-progress (**Closing – Opening**) stocks. All figures are in £000. $\hat{B}$ = estimated bias.

The leading method for bias reduction with voluntary samples is *poststratification* (for example, Holt & Smith 1979, Jagers 1986, Smith 1991, Little 1993). Taking for simplicity the case of a single outcome of interest, two ingredients are required for this method: (i) a list, preferably (close to) exhaustive, of covariates likely to be (highly) correlated with the outcome; and (ii) the ability to gather data on these covariates both in the voluntary sample and in the population itself. Dividing each covariate into strata and cross-tabulating the resulting categorical variables, poststratification involves (a) estimating both population and voluntary sample prevalences in the cells of this stratification grid, and (b) re-weighting the voluntary sample to match the estimated population prevalences. Ideally the stability of this method should be checked by *sensitivity analysis* (see Draper *et al.* 1993a for examples), varying the covariates used and the cut-points defining their strata across plausible ranges and seeing whether the bias-adjusted estimates are similar. The (approximate) success of this method rests on the assumption that (most or all of) the important covariates have been correctly identified, measured, and adjusted for.

Variable	Probability Sample			Voluntary Sample		
	Industry 1	Industry 2	Industry 3	Industry 1	Industry 2	Industry 3
Register employment	152	153	197	334	276	605
Register turnover	11,949	7,171	9,775	25,206	16,425	28,388

Table 4.4 Comparison of probability and voluntary samples on median register employment (numbers of people) and turnover (£000), by industry, in the first quarter of 1997 (results for the other two quarters were similar).

In this example the only available covariates are register employment (*E*) and turnover (*T*), which are fairly highly correlated in both the *P* and *V* samples (for example, the correlation, with both variables on the log scale, is +0.74 in the voluntary sample). Table 4.4 shows that at least some of the discrepancy between the probability and voluntary samples should indeed be explainable on the basis of *E* and/or *T*: the 30% of the quarterly probability sample that chose to volunteer monthly data heavily over-represented large companies.

To avoid redundancy we present poststratification results here only for one industry (results were similar with the other two industries). With only 17 companies per quarter in this industry in the voluntary sample, bivariate stratification on both *E* and *T* would leave empty cells, which does not permit re-weighting, so in the work presented here we first stratified only on register turnover (in any case the high correlation between *E* and *T* indicates that there is not much information in *E* after *T* has been accounted for). We chose four strata, with the smallest cutpoint selected so that the lowest stratum had at least one company in both samples, and with the other two cutpoints chosen to spread the rest of the distribution out approximately evenly.

Table 4.5 indicates how the probability and voluntary samples in industry 1 were distributed across strata based on register turnover. This provides another view of how sharply the large companies were over-sampled in the voluntary survey, for example, 43% of the probability-sampled companies were in the smallest register-turnover stratum, versus 6% in the voluntary sample. The weights used in the poststratification are also given in this table; for example, the voluntary-sample company in the lowest stratum was given weight $(30/70)/(1/17) \cong 7.29$, whereas the 6 voluntary companies in the highest stratum received weight $(14/70)/(6/17) \cong 0.57$.

Register turnover intervals (£000)	*P*	*V*	Weight
[0-8,455]	30	1	7.29
(8,455-14,784]	12	4	0.73
(14,784-84,657]	14	6	0.57
(84,657-2,284,224]	14	6	0.57
Total	70	17	–

Table 4.5 Frequency distribution of probability (*P*) and voluntary (*V*) samples, across the four register turnover strata, together with the poststratification weights.

Table 4.6 presents the results of the bias reduction arising from poststratification on register turnover. Separately for each of the stocks categories {Opening, Closing, and (Closing − Opening)}, the *P* column gives the probability-sample estimate (reported previously in Table 4.1-Table 4.3), the *PV* column is the voluntary-sample estimate re-weighted by the poststratification on register turnover, $\hat{B} = PV - P$ is the estimated bias after poststratification, and $\hat{R}$ is the percentage (relative) reduction in estimated bias yielded by the poststratification. For example, in 1997/Q2 the raw voluntary-sample estimate for industry 1 was 6,148, giving an estimated bias of +2,692 (Table 4.1); after re-weighting the new voluntary-sample estimate is 3,556, with an estimated bias of +100 (Table 4.6); and diminishing the estimated bias from 2,692 to 100 represents an estimated bias reduction of $(2{,}692 - 100)/2692 \cong 96.3\%$. Poststratification has resulted in massive estimated bias reductions ranging from 81% to 97% for the opening and closing stocks, but has produced a more modest estimated improvement in the crucial difference (Closing − Opening), with gains from 53% to 90%.

Period	Opening				Closing			
	P	PV	$\hat{B}$	$\hat{R}$ (%)	P	PV	$\hat{B}$	$\hat{R}$ (%)
Q1	3,420	3,223	-197	90.2	3,456	3,556	+100	96.3
Q2	3,456	3,556	+100	96.3	3,455	3,541	+86	96.6
Q3	3,455	3,541	+86	96.6	3,898	4,628	+730	81.4

Period	Closing – Opening			
	P	PV	$\hat{B}$	$\hat{R}$ (%)
Q1	36	333	+297	56.8
Q2	-1	-15	-14	89.9
Q3	443	1,087	+644	53.2

Table 4.6 Results, by period, from poststratifying on register **turnover**. In each of the stocks categories {Opening, Closing, and (Closing − Opening)}, P is the probability-sample estimate, PV is the poststratified voluntary sample estimate, $\hat{B} = PV - P$ is the estimated bias after poststratification, and $\hat{R}$ is the percentage reduction in estimated bias arising from the poststratification.

Period	Opening				Closing			
	P	PV	$\hat{B}$	$\hat{R}$ (%)	P	PV	$\hat{B}$	$\hat{R}$ (%)
Q1	3,420	3,301	-119	94.1	3,456	3,598	+142	94.7
Q2	3,456	3,598	+142	94.7	3,455	3,549	+94	96.3
Q3	3,455	3,549	+94	96.3	3,898	4,528	+630	84.0

Period	Closing – Opening			
	P	PV	$\hat{B}$	$\hat{R}$ (%)
Q1	36	307	+271	60.6
Q2	-1	-49	-48	65.5
Q3	443	979	+536	61.1

Table 4.7 Results, by period, from poststratifying on register **employment**. In each of the stocks categories {Opening, Closing, and (Closing – Opening)}, P is the probability-sample

estimate, PV is the poststratified voluntary sample estimate, $\hat{B} = PV - P$ is the estimated bias after poststratification, and $\hat{R}$ is the percentage reduction in estimated bias arising from the poststratification.

Sensitivity analysis on the poststratification process is straightforward. For example, basing the strata on register employment and using three strata instead of four (with stratum definitions [20-215], (215-449], and (449-12,378]), chosen to create approximately equal-sized groups in the voluntary sample, yielded the results in Table 4.7. The two approaches to poststratification have in this case led to similar amounts of bias reduction, although this need not always be true. In practice, when a "gold-standard" (such as the probability-sample results here) is not available, any differences revealed by a comparison of this type may indicate that other variables should ideally have been part of the stratum definitions, that is that poststratification may not have been entirely successful in removing the selection bias present in the voluntary sample.

4.3 Quota sampling

For a straightforward definition of quota sampling we turn to Särndal *et al.* (1992, p 530)

> "Quota sampling is often used in market research. The basic principle is that the sample contains a fixed number of elements in specified population cells. Suppose that the population is divided according to three controls: sex, age group, and geographic area. With two sexes, four age groups, and six areas, we get a total of $2 \times 4 \times 6 = 48$ population cells. In each cell, the investigator fixes a number (a "quota") of elements to be included in the sample. Now the interviewer simply "fills the quotas", that is, interviews the predetermined number of persons in each of the quota cells. These may be the first persons encountered, or it may be left to the interviewer to exercise judgement in the quota selection. *The method resembles stratification, but the selection within strata is non-probabilistic* [emphasis added]. Because that selection is non-probabilistic, there is neither an unbiased point estimate nor a valid variance estimate within the cell."

(Also see Deville 1991 for one attempt at establishing a theoretical basis for quota sampling.) In practice quota samplers often simply assume that the population units which end up in each of the cells are like what one would have obtained with simple random sampling within each cell, both for want of anything better to assume and because this assumption turns quota sampling into stratified random sampling (StRS) and the usual estimates of error (for example, Cochran 1977) are then available. Indeed, as Särndal *et al.* (1992) note, adopting a model-based approach in which the y_j are assumed to be random variables with $\mathrm{E}_\xi(y_j) = \mu_h$, $\mathrm{V}_\xi(y_j) = \sigma_h^2$, where h indexes the cell in the quota-sampling grid in which y_j is observed, yields precisely the same estimate of the population total t as with StRS,

$$\hat{t} = \sum_{h=1}^{H} N_h \bar{y}_{sh} \qquad (4.2)$$

where H is the number of cells in the grid and N_h and $\bar{y}_{sh}$ are the population size and sample mean in cell h, respectively. Moreover, the usual StRS estimated variance of this estimator,

$$V_\xi(\hat{t}) = \sum_{\mathrm{h=1}}^{\mathrm{H}} N_h^2 \frac{1-f_h}{n_h}\left[\frac{1}{n_h-1}\sum_{sh}(y_j - \bar{y}_{sh})^2\right] \tag{4.3}$$

where n_h is the sample size in cell *h* and $f_h = \frac{n_h}{N_h}$, is unbiased under this model. Thus valid interval estimates for such quantities as the population total or mean and stratum (cell) means are available, *under the assumption that the model is correct* (see also chapter 9). In Bayesian treatments of sample surveys this sort of assumption would be described as a judgement that the sampled and unsampled units in each of the population cells are *exchangeable* (see Draper *et al*. 1993b for discussion), which just means that one's predictive uncertainty for both the sampled and unsampled units before any data are gathered would be the same.

If additional relevant stratifiers (what Särndal *et al*. (1992) called *controls* in the quote above) are available in the quota sample and population prevalences are known, poststratification (as in Section 4.2) within each cell can be employed to adjust for possible selection bias arising from the haphazard choice mechanism (see Smith 1983, 1993 for examples).

Quota sampling does not seem to be much in use in European structural and short-term business surveys at present, although a kind of quota sampling that could also be termed judgemental sampling (see Section 4.4 below) is employed by many EU Member States in the compilation of price statistics (Eurostat 1998:07).

4.4 Judgemental sampling

As noted by Eurostat (1996:04), "several [EU] countries use judicious samples based on a high coverage of relevant characteristics (for example, production, employment, and turnover). This mainly concerns production and output price indices for which there is no register of products." In effect, such samples are based on expert judgement as to representativeness rather than full probability sampling.

An example of how this may arise occurs in one or more stages of the sampling process supporting the creation of producer price indices. For instance, Eurostat (1998:07, abbreviated E98) contains an extensive discussion of methodological aspects of estimating producer prices on the export market; most of the material in this section is based on this document.

4.4.1 Producer price index construction in the EU

Background on the problem addressed by export-market producer price indices is as follows.

> "Producer price indices in general should cover the prices of all commodities produced in a given country in order to be consistent with [the country's overall index of production]. ... While total producer price indices (PPI) show the evolution of prices for goods *produced* on the domestic market, irrespective of whether they are sold on the domestic market or abroad, producer prices on the export market (PPI$_x$) only take into account the prices for those commodities which are sold abroad. ... The main purpose of the PPI$_x$ is to provide rapid information on business cycle movements, that is, to serve as an economic

> indicator. Furthermore PPI$_x$ also serves as a deflator for foreign trade data and for national accounts. ...
>
> [The] PPI$_x$ for a given industry group should be calculated as a weighted average of commodity price indices, based on a sample of enterprises and samples of representative commodities. Thus the first step in the compilation of a PPI$_x$ is the selection of a basket of representative "goods," that is, headings at a given level of a product nomenclature (such as PRODCOM or HS). In accordance with the selected goods, enterprises have to be chosen which produce these goods on a regular basis destined to be sold abroad. The last step consists in defining in each enterprise the products representing these goods, for which prices will then be reported each month." [E98, pp. 4–5]

In other words, the creation of a PPI$_x$ typically involves three stages of sampling: (i) choosing a kind of market-basket of goods, (ii) selecting enterprises (companies) making those goods, and (iii) taking a sample of actual products representing the goods made by the enterprises. In practice each stage of selection in this hierarchy may use one or more sampling methods in a more or less formal way, for example, stratification, probability proportional to size, cut-off sampling, and/or expert judgement. Here are two examples from specific EU Member States:

1. In the Netherlands, "The selection of products and reporting units is based on detailed base year production and consumption data from different statistical sources, such as production statistics and foreign trade statistics. ... In order to guarantee a minimum quality of price indices, the following rule applies: per commodity group the selected reporting unit should on average cover 80% of sales (cut-off method). If for a particular commodity more than 25 reporting units are required in order to attain 80% coverage, a random sampling method can be applied. ... Once the reporting units have been selected, the next step is to select for each reporting unit certain products within a specified commodity group. The price statistician knows for what kinds of products he wants to gather prices from the reporting unit. So, with the help of a field surveyor, a visit is made to the reporting unit. The reporting unit is asked to specify the price of a product, within the commodity group, which is representative for the export. At least one, but normally two or more, prices are asked for. ... At present about 7,000 export price quotations are collected at frequent intervals from about 5,500 reporting units." [E98, pp. 23–24]

2. In Sweden, "The sample of representative items is revised annually and is made in four steps: (i) Industrial activities (as specified by [the Swedish version of] SIC92) are sampled by cut-off according to export value. Within each activity (ii) commodities (as specified by HS) are then also sampled by cut-off according to Foreign Trade Statistics which have been processed for the national accounts. (iii) Producers of selected commodities are then sampled by cut-off from the Foreign Trade Statistics register of exporters. (iv) Finally, representative items are selected [judgementally] in consultation with the respondent (producer). They are selected with preference to products with high sales values, which could be expected to be sold during all months, and if possible are representative of price movements within the commodity group (HS number)." [E98, p. 44]

As these excerpts demonstrate, the choice of detailed commodity specifications is likely to involve discussion with each enterprise as a basis for expert judgement. The Swedish example shows that these commodities are typically chosen to be representative of price changes, and to be sold both frequently (so that monthly data are available) and for a long period of time. It is important to assess the accuracy of the types of samples just mentioned. For example, if products are chosen because they have enjoyed frequent sales, this may be due to low prices, and those prices may, during periods of rising inflation, increase more than others.

It does not appear that many EU Member States are attempting at present to assess the bias or sampling variability with which their PPI$_x$ are estimated. The effects of judgemental sampling are normally difficult to quantify, but there are several approaches which can be adopted, some of which rely on the existence of other information, and some of which are only available through additional studies. We conclude this section with a discussion of some methods currently in use in the UK.

4.4.2 The UK experience

The first point to note, in the context of price indices, is that there is rarely a frame with product information from which commodities can be selected. As mentioned above, this means that sampling is usually restricted to choosing an enterprise, and then identifying a "representative" product on a judgemental basis. There has been a tendency in the UK PPI to obtain more than one quote from businesses for similar products, which in practice gives little additional information, since businesses usually have consistent pricing policies; it would be better to obtain quotes for different products, or to sample a new business. This is especially important if the sample size in terms of number of price quotes is fixed or constrained.

Small-scale studies of the effect of this sampling can be made by enumerating the products manufactured by a business, selecting a probability-based sample, and then looking at the price movements over a short period in comparison to the existing judgemental sample. This approach is expensive in collecting additional information and forming the product list to sample from.

The UK is in the process of transition from a judgemental sample to a sample based on this concept. Lists of product sales at the detailed (8-digit) level of the PRODCOM classification are obtained as part of the PRODCOM survey for a (probability) sample of businesses from the IDBR. These will then be used to form a frame from which sampling of 8-digit products can take place according to a probability mechanism in the PPI, giving a two-phase design. There is still an issue of which product to choose within an 8-digit heading, but at least the business-product pair will be selected by a probability mechanism from the PRODCOM sampling, and appropriate weighting can be used to give a design-unbiased estimator of the "population PPI". The first stage in the introduction of this design is underway in the UK, and results comparing the current judgemental system (which also inherits many characteristics of a previous voluntary survey) and the new probability-based system are expected around April 1999.

There are particular problems with the products of some industries which may make judgemental selection of a "representative" product extremely difficult. In the clothing industry, for instance, items and fashions change on a seasonal basis, and getting a continuous price quote for a transient line is impossible. Thus there will be a tendency to select continuously-produced products, even when these do not accurately represent the overall price movement under the appropriate heading.

In a similar way it might be expected that "typical" rather than representative products are identified, and that for this reason minority production (which might have a more volatile price) may be missed. This is very difficult to assess: the information required is about the proportion of extreme price movements, which requires a large sample for estimation. However, in cases where product identification instructions draw attention to this problem, it should be noted as part of the quality assessment that this may be an issue.

Some assessment of the quality of a judgemental sample can also be made using the model-based approach by invoking the ignorable sampling assumption (see Chapters 2 and 9). If we assume (probably falsely) that the judgemental sample is approximately representative, then we can calculate the variability of prices in product categories (choosing a higher or lower level depending on the sample size available so as to obtain a reasonable estimate). This helps to assess the "sampling variability" of the judgemental sample, and by reallocating the sample using a Neyman-type allocation and calculating the expected variance (noting that the expected variance is smaller than what will be achieved in practice because it uses the same data for allocation and sampling variance estimation), the two can be compared. This approach has been adopted in the optimisation of the UK CPI, where − for example − the number of quotes for potatoes was increased because of the variability induced by the high price of imported new potatoes at certain times of the year.

4.5 Cut-off sampling

Once again Särndal *et al*. (1992) is a good source for a simple description of *cut-off sampling*. As in Section 4.1 let the *N* units in the population *U* be indexed by *j*, and define π_j as the probability that unit *j* is chosen in the sample.

> "Probability sampling requires that $\pi_j > 0$ for all $j \in U$. There are sampling methods in current use that employ probability selection with $\pi_j > 0$ for part of the population *U*, whereas $\pi_j = 0$ for the remainder of *U*. Such methods take an intermediate position between probability sampling and non-probabilistic selection with π_j that are unknown throughout the population. One of these techniques is cut-off sampling. In cut-off sampling there is a usually deliberate exclusion of part of the target population from sample selection. This procedure, which leads to biased estimates, is justified by the following argument: (i) that it would cost too much, in relation to a small gain in accuracy, to construct and maintain a reliable frame for the entire population; and (ii) that the bias caused by the cut-off is deemed negligible. In

> particular, the procedure is used when the distribution of the values $y_1, \ldots, y_N$ is highly skewed, and no reliable frame exists for the small elements. Such populations are often found in business surveys. A considerable portion of the population may consist of small business enterprises whose contribution to the total of a variable of interest (for example, sales) is modest or negligible. At the other extreme, such a population often contains some giant enterprises whose inclusion in the sample is virtually mandatory in order not to risk large error in an estimated total. One may decide in such a case to cut off (exclude from the frame, thus from sample selection) the enterprises with few employees, say five or less. The procedure is not recommended if a good frame for the whole population can be constructed without excessive cost."

(See Sugden & Smith (1984) and Haan, Opperdoes & Schut (1997) for more on cut-off sampling.)

As an illustration of the kind of data for which cut-off sampling might be used, consider the annual UK *Annual Business Inquiry* (ABI) survey, which estimates current employment, turnover, and value added based on a sample chosen with the aid of register employment and turnover (Table 4.8; the register contains information from 3-6 months before the survey). ABI stratifies on industry (by 3-digit SIC), region (12 categories) and register employment, over-sampling large companies (compare the raw-mean and weighted-mean columns in Table 4.8 to see how sharp the over-sampling is). The sample weights required to compensate for this varied in 1996 from 1 to 27.9 with a mean of 3.45. We can use the samples of size 2,737 and 2,453 in 1995/96 as the basis of an exercise in which (a) simulated populations are created and (b) cut-off samples are chosen from these populations, to explore the biases that result from ignoring or modelling the smallest companies.

Variable	Raw mean	Weighted mean
Register employment	211.1	79.9
Returned employment	195.4	76.9
Register turnover (£000)	33,491.4	11,307.1
Returned turnover (£000)	31,374.6	10,757.5

Table 4.8 Variables available in the analysis of the UK ABI survey presented here (values are from the 1996 sample).

Returning to the quote from Särndal *et al.* (1992),

> "Let U_c denote the cut-off portion of the population and let U_0 be the rest of the population, from which we assume that a probability sample is selected in the normal way. The whole population is thus $U = U_0 \cup U_c$. Each element in the cut-off portion has zero inclusion probability; that is, $\pi_j = 0$ for all $j \in U_c$. Let $\hat{t}_0$ be an

> estimator of $t_0 = \sum_{U_0} y_j$, for example, $\hat{t}_0 = \sum_{s_0} \frac{y_j}{\pi_j}$. But we need an estimator of the whole total $t = \sum_U y_j$. How can this be achieved?"

The two possible courses of action in this situation are evidently to ignore the cut-off units altogether or to try to estimate their contribution to the total. In the next two subsections we consider each of these possibilities in turn.

4.5.1 Variation 1: Ignore the cut-off units

As Särndal *et al.* (1992) note, in this variation, which is equivalent to estimating the total across the cut-off units as zero,

> "The statistician may be willing to assume that $T_c = \sum_{U_c} y_j$ is a negligible portion of the whole total $t = \sum_U y_j$. If $\hat{t}_0$ by itself is used to estimate t, the relative bias is
>
> $$\frac{\mathrm{E}(\hat{t}_0) - t}{t} = -\frac{t_c}{t} = -\left(1 - \frac{t_0}{t}\right), \tag{4.4}$$
>
> which is negative but negligible under the assumption. We assume that y is an always positive variable."

Continuing the ABI example above, consider a given industry, with an outcome variable such as turnover, and using a proxy variable for turnover such as number of employees. One way to define the cut-off units U_c is by (a) sorting all companies in the register on employee numbers, obtaining $x_{(1)},\ldots,x_{(N)}$, where $x_{(j)}$ is the *j*th smallest number of employees; (b) calculating the cumulative sum of employee numbers from the smallest to the largest companies, obtaining $S_1 = x_{(1)},\ldots,S_J = \sum_{k=1}^{j} x_{(k)},\ldots,S_N = \sum_{k=1}^{N} x_{(k)}$; and (c) cutting off all the companies for $S_j \le (1-\varepsilon)S_N$, for some small ε such as 0.05. Here it is as though the population of interest is defined to be just the top $100(1-\varepsilon)\%$ companies in employee numbers. Probability sampling from the resulting set U_0 of non-cut-off companies could now be undertaken, as Särndal *et al.* (1992) mention, or complete enumeration of the y values in U_0 could occur.

A strategy related to the one just outlined would be to simply define the population of interest to be all companies with (say) 5 or more employees, sample from the companies with (say) 5-200 employees, and attempt a full enumeration of the companies with more than 200 employees. Here one point of ignoring the tiny companies by definition is that laws preventing the governmental survey burden on small companies from being too great may make it impractical or impossible to get data from them in any case. However, by choosing ε appropriately and over-sampling with sufficient vigour on the largest companies (defined by employee numbers), this approach is seen to be a close approximation of the method in the previous paragraph, on which we focus below.

To estimate the bias arising from variation 1 of cut-off sampling, for each of several values of ε we repeatedly (100 times) (a) drew a sample of size 2,453 (the 1996 ABI sample size) with replacement from the ABI data but with unequal selection probabilities determined by the sampling weights, to create a pseudo-population reflecting the actual distribution of UK companies (this is a kind of *weighted bootstrap*; see Efron & Tibshirani 1993), (b) used the register employment variable in this population to cut off the lower 100ε% of the companies (by cumulative employee numbers, as described above), and (c) estimated the total returned turnover by the total across the companies not cut off. To focus on bias issues we are thus employing the strategy of full enumeration within U_0.

ε	Relative bias, in % (SE)	Maximum bias, in %	Average employment of businesses cut-off (SE)	% of businesses cut-off
0.20	-12.5 (0.16)	-16.6	54.0 (0.6)	75.8 (0.2)
0.15	-9.15 (0.12)	-12.1	36.1 (0.3)	66.7 (0.2)
0.10	-5.99 (0.08)	-7.8	24.5 (0.1)	53.1 (0.2)
0.05	-2.98 (0.04)	-3.9	15.7 (0.1)	32.5 (0.2)

Table 4.9 Simulation results from cut-off sampling the 1996 ABI data, based on 100 simulation repetitions (SE = Monte Carlo standard errors). The sample size in each case was 2,453.

Table 4.9 presents a summary of this simulation exercise. (Results with larger sample sizes of 5,000 and 10,000 were virtually identical.) To interpret the results in the table, consider the row for $\varepsilon = 0.20$ (that is, using a 20% cut-off). Across the 100 simulation replications, the average amount by which the cut-off estimate fell short of the total across all 2,453 companies was 12.5% of the true total, and the maximum such relative bias across the 100 replications was 16.6%. On average the cut-off companies had about 54 employees or less, and such companies made up about 76% of all companies. It can be seen from the $\varepsilon = 0.05$ row in the table that, with data of this type, cutting off the 5% smallest companies (in terms of total employees in the register) leads to a downward bias of about 3% in total turnover, while allowing the sampling process to ignore about a third of the companies. Whether a bias of this magnitude is acceptable depends on the context.

In practice the success of this variation of cut-off sampling varies strongly with ε, in a population- and problem-specific manner. For instance, the discussion thus far has emphasised the estimation of the *level* of, for example, turnover at one point in time rather than the *change* in turnover level over time. When the main aim is to estimate change, the proportion of cut-off units in the population may be taken to be higher (for a given bias tolerance) than in the case of a level, because some of the bias should cancel in the subtraction underlying the change estimate. To illustrate this point, we replicated the analysis of Table 4.9 on both the 1995 and 1996 ABI samples, repeatedly (100 times) creating pseudo-populations for each year and recording the absolute and relative biases from ignoring the cut-off units in 1995, 1996, and the change from 1995 to 1996.

Table 4.10 presents the results of this second simulation. Columns 6 and 7 (counting from the left) in the table exhibit the expected bias cancellation, in absolute and relative terms, in estimating the change from 1995 to 1996; for example, at $\varepsilon = 0.15$, biases of 7-9% in the individual years are reduced to 5% when the change from year to year is the quantity of principal interest.

ε	Bias 1995		Bias 1996		Bias (1996 – 1995)	
	Absolute	Relative (%)	Absolute	Relative (%)	Absolute	Relative (%)
0.20	-4,268	-10.0	-3,253	-12.3	1,014	-6.3
0.15	-3,114	-7.3	-2,373	-9.0	741	-4.6
0.10	-1,982	-4.7	-1,558	-5.9	424	-2.6
0.05	-981	-2.3	-776	-2.9	205	-1.3

Table 4.10 Absolute (in £M) and relative (in %) bias results from ignoring the cut-off units in estimating the 1995 and 1996 total turnover values, and the change from 1995 to 1996, in the UK ABI survey. The 1996 results differ a bit from those in Table 4.9 because a different random number seed was used in each case.

4.5.2 Variation 2: Model the cut-off units

The other leading approach to estimating population totals with cut-off sampling is to try to estimate the contribution to the total provided by the cut-off portion of the population U_c. As Särndal *et al.* (1992) put it,

> “A second approach is to use a ratio adjustment for the cut-off. Let x be an auxiliary variable, for example, the variable of interest measured for the entire population at an earlier date, or some other known variable roughly proportional to the current variable of interest y. Let
>
> $$R_{U_0} = \frac{\sum_{U_0} y_j}{\sum_{U_0} x_j} \tag{4.5}$$
>
> and let
>
> $$\hat{R}_{U_0} = \frac{\sum_{S_0} \frac{y_j}{\pi_j}}{\sum_{S_0} \frac{x_j}{\pi_j}} \tag{4.6}$$

> be the [design]-consistent estimator of R_{U_0}, based on the probability sample from U_0. To extend the conclusions to the whole population, an unverifiable assumption is necessary. Assume that $R_{U_0} = R_U = \frac{\sum_U y_j}{\sum_U x_j}$. Then $\hat{R}_{U_0}$ can serve to estimate R_U as well, and by ratio adjustment we arrive at
>
> $$\hat{t}_{cut-off} = \left(\sum_U x_j \right) \hat{R}_{U_0} \tag{4.7}$$
>
> as an estimator of the whole current total $t = \sum_U y_j$, assuming $\sum_U x_j$ or a close estimate of it is available. The relative bias is approximately
>
> $$\frac{\mathrm{E}\left(\hat{t}_{cut-off}\right) - t}{t} = \frac{R_{U_0}}{R_U} - 1 \ , \tag{4.8}$$
>
> which can be positive or negative. It is zero if the assumption $R_U = R_{U_0}$ holds. This assumption is one that the statistician may be more inclined to make than the assumption in the first approach that t_c / t is negligible."

This strategy is based on ratio estimation, but this is not the only option: ratio estimation is equivalent to regression estimation with the presumed regression line going through the origin (see Cochran 1977), and one may use regression estimation without the intercept being thus restricted. Moreover, as we will see in Chapter 9, the regression estimation could occur either on the raw scale, for both the *x* and *y* variables, or on the log scale.

Eurostat (1997:04) contains another example of Variation 2: "When units are selected with certainty following a structural auxiliary variable, such as yearly value added, a more sophisticated indicator could be built using an econometric model in order to estimate the effect of enterprises not selected." The approach in this variation is now taken in most or all EU regulations involving cut-off sampling (Eurostat 1997:06, 1997:07). Because of its dependence on modelling assumptions we postpone further discussion of this variation to Chapter 9.

In both variations, one problem is that legislation may say one should obtain data from the companies providing the top (say) 95% of *current* employment, but in fact *past* employment is typically used (whatever is the most current figure, which may be anywhere from 3-6 months to 1-2 years out of date, depending on EU Member State) as a proxy. The seriousness of this problem naturally grows with the gap in time between current and register employment.

4.6 Conclusions

We conclude this chapter with a set of recommendations for each of the non-probability-sampling situations examined in the sections above.

- **Recommendations**: Model reporting in business surveys involving *voluntary sampling* should
    - ➢ Acknowledge explicitly that voluntary sampling has been used; and
    - ➢ Present estimates and uncertainty assessments both with and without poststratification on the most important available covariates, so that consumers of the analysis can see both (a) whether they agree that all relevant covariates have been accounted for and (b) the direction and magnitude of the bias adjustment.
- **Recommendations**: Model reporting in business surveys involving *quota sampling* should
    - ➢ Acknowledge explicitly that quota sampling has been used;
    - ➢ Present provisional estimates and uncertainty assessments as if the data had been gathered using stratified random sampling, with the same stratification grid as that used to define the quotas; and
    - ➢ Present evidence, if available, demonstrating that the quota samples within the cells of the grid provide approximately unbiased estimates of the population means in those cells. This evidence could take the form of sensitivity analyses showing that the results of principal interest are little changed when stratification with respect to additional plausibly relevant variables is undertaken. If no such evidence is available, the quota sampling estimates and uncertainty assessments should be presented with an explicit statement that the unbiasedness of the estimated cell means has not been conclusively established.
- **Recommendations**: Model reporting in business surveys involving *judgemental sampling*, for example, in the creation of producer price indices, should
    - ➢ Routinely seek and present evidence that judgementally "typical" products are in fact representative of actual price movements, and
    - ➢ Periodically calculate the variability of prices in product categories based on an assumption that the judgemental sample is approximately representative.
- **Recommendations**: Model reporting in business surveys involving *cut-off sampling* without any attempt to estimate the contribution of the cut-off population units (variation 1 in Section 4.5.1) should
    - ➢ Provide evidence, of a simulation nature or otherwise, that the percentage of population units cut off and ignored leads to acceptably low bias with problems and populations similar to those currently under study.

Part 2: Non-sampling errors

5 Frame errors

Eva Elvers[3], Statistics Sweden

5.1 Introduction

Among the non-sampling errors that contribute to the overall inaccuracy are *frame errors*, to be described in this chapter. The construction of a frame is one of the first steps in the production process and essential for the steps to follow. The frame must, of course, be defined with regard to the final goal, the resulting statistics. These are estimates of finite population parameters (FPPs). Ingredients in such parameters are

- statistical measure (total, mean, median, etc);
- variable (production, number of hours worked, etc);
- unit (enterprise, kind-of-activity unit, etc);
- domain (sub-population, for example defined by a standard classification like NACE Rev. 1);
- reference times; both units and variable values relate to specific times.

The reference times are mostly time intervals, like a calendar year, a quarter, or a month. However, some variables may refer to a point in time, for example the starting point of the period. Usually reference times agree for all variables and units in a FPP. This means , for example, for monthly statistics that the delineation of units should refer to the current month. It follows from the above that units, classifications, other auxiliary variables, and reference times are essential to statistics – and so also to the frame.

The emphasis here is to be on the assessment of quality, but some background is necessary. Section 5.2 deals with a Business Register and its use as frame – a foundation without which the statistics can hardly be built. Section 5.3 describes frame and target populations. The accuracy to be measured depends on the frame but also on estimation procedures; Section 5.4 describes some situations. Sections 5.5-5.7 illustrate; showing administrative sources, time delays and frame construction, and frame differences and quality assessment measures. There are some summarising conclusions in Section 5.8.

5.2 A Business Register and its use as a frame

5.2.1 Units, delineation, and variables

The abbreviation SIC will be used for convenience for Standard Industrial Classification, meaning NACE Rev. 1 and often referring to the primary activity.

[3] Many persons have contributed with data, examples, and comments, especially Pär Lundqvist at Statistics Sweden, and Ole Black, John Perry, Ian Richardson, and Mark Williams at ONS, UK.

A Business Register is here – in agreement with the Council regulation No 2186/93 on drawing up business registers for statistical purposes, and, hence, also in agreement with the Council regulation No 696/93 on statistical units – regarded as a database with

- a set of units; at least enterprise, legal unit, and local unit;
- a set of variables to each unit; such as SIC code and size, for example the number of employees;
- a set of time stamps (explicit or implicit); at least the time of registration for updates;
- links between units, with time stamps.

The BR builds on administrative information, investigations of its own, and information from statistical surveys. Note that survey feedback has to be used with care when sampling with co-ordination over time in order not to distort the randomness of the sample, see Ohlsson (1995). The information in the BR is as recent as possible. This goes both for each variable and for the delineation of units. The delineation refers not only to single units but also to information on links between units, for example links between legal units and enterprises. How recent the BR information is varies between variables and also between units, depending on updating procedures.

The BR shows each unit with its SIC code, size measures, links to other units etc. Variables on a higher level in the hierarchy of units are in many cases derived by aggregation from a lower level, for example number of employees and SIC code. Some variables may, however, not be available on a low level, for example turnover connected to VAT (value-added tax).

The choice of which variables to put on the BR should consider both the unit level and the usefulness as auxiliary variables in different procedures (for updating, creating frames, estimation etc).

5.2.2 Updating the BR using several sources

Some typical examples of updates are as follows. Information on births and deaths arrives from administrative sources regularly with known frequency. The time-lag between an actual event and when it is recorded may be different for births and deaths. For example, the time-lag from the first paying of VAT to birth in the BR may be short in comparison with the time-lag from ceased activity to registration of death in the BR if that is based on a de-registration at fiscal authorities. A survey may detect the no-activity state much quicker than the BR – the difficulty for the survey may be to distinguish between this state and nonresponse.

The information available on an enterprise at its birth in the BR may be fairly limited, and it usually takes some time before it has an adequate size and SIC code. For certain statistics, for example on investments in fixed assets (investments for short in the following), an early detection of new activity is important. At the time the investment is made, there are likely to be few employees and the unit may not yet have turnover. Hence, it is desirable to find additional sources of information which show such activities at an early stage. It is important that these sources are consistent over time and space.

The sources of the BR for births, deaths, and updates could be PAYE (this abbreviation will be used in the following for administrative information from the collection of taxes on earnings, which includes employment) and VAT. The BR may have a survey of its own, for example concerning units, links between units, and classifications.

When an update is made, not only a change of the value is made, but there is also a notation as to time. The simplest thing is to note the time of *registration*. There should preferably be also a time of *occurrence*. A new SIC code may for example be registered in February 1998 but be valid from January 1996. The time is possibly known implicitly from the source. Time stamps add to the information and they are valuable in demographic studies, but they also make the handling more complex.

The use of several sources makes it necessary to have some identification. There may, for example, be an identification number (id.nr) for legal units used by fiscal authorities, that is, the BR obtains VAT data by legal unit id.nr.

Some identification is necessary not only to update but also to merge information from different sources. Such merging is simple if there is a unique identification number common to all sources. This is, however, rarely the case. For example, there are different id.nrs in Germany and Ireland for the two administrative data sets regarding VAT and PAYE, making it necessary to merge the information by name and address.

In Sweden, there is a singe number for a legal unit, but an enterprise consisting of several legal units may choose to report VAT and PAYE data for one and the same activity as belonging to different legal units. This means that the legal unit numbers are not identification numbers in the sense of business activity.

The UK experience is that it has found business structures to be complex and based on administrative procedures that are not always suitable for statistical inquiries. The VAT unit is there to facilitate the collection of VAT, and it may not be able to provide the survey information required. Also some employers maintain separate PAYE systems for salaried and non-salaried workers, giving two administrative units and making it necessary to merge information from the two systems when updating the frame.

The above examples show that duplicates can easily arise on the BR – unless counter-actions are taken – since a single activity may lead to several births through different administrative sources.

5.2.3 The BR as a frame – units, variables and reference times

Consider first units for different purposes in different parts of the production process. Sampling is performed in one or possibly more stages with a sampling unit at each stage (for example a single stage with enterprise as the sampling unit). The data collection is addressed to the *reporting unit* (for example the enterprise through a questionnaire) or more generally to the source of information (which could be an administrative register). The observations of the statistical survey are tied to the *observation units*. The reporting unit can be equal to the observation unit or be different: an enterprise as the reporting unit and a kind-of-activity unit as the observation unit provides an example of the latter case.

Note: The terminology is not unique; *collection unit* is sometimes used for reporting unit, and *reporting unit* is sometimes used for observation unit.

It is important to consider the domains of estimation when choosing units. The observation unit should not cut across several domains; for example an enterprise consisting of several kind-of-activity units should not be the observation unit for statistics that are based on kind-of-activity units, so-called functional statistics.

Here, the emphasis is on Structural Business Statistics (SBS) and Short-Term Statistics (STS), with units of the FPPs being: enterprise for SBS, enterprise for parts of the STS, kind-of-activity unit (KAU) for parts of the STS, and then possibly also legal unit, local unit, and local kind-of-activity unit.

The BR (as defined here) is such that there is an agreement between the register units and the units to be used in business statistics. The step from the BR and its units to a frame population is then principally short and simple. It involves making a list of units with regard to SIC code and possibly also size; variables that are available in the BR. The most pronounced principal difficulty may be the kind-of-activity unit, depending on whether it is included in the BR or not. This unit could alternatively be created at the data collection stage (KAU from enterprise, and local KAU from local unit).

Struijs & Willeboordse (1995) discuss units and changes of units.

5.3 Frame and target populations

5.3.1 Target population

As stated, the target parameters have the reference time for both units and variables equal to the current month/quarter/year. The target population could for example be all enterprises or all kind-of-activity units in the manufacturing industry which are active in the current period.

5.3.2 Frame, and frame population

Ideally there is a perfect frame which lists every unit in the target population once and only once together with basic design variables. In reality the frame is affected by various imperfections for several reasons, for example time delays and coding mistakes. For business statistics, like SBS and STS, the frame is normally based on a BR.

The frame population for a particular survey is based on the target population of that survey. It is normally expressed in the same way as the target population, that is, in terms of units, SIC codes, and possibly size; for example “all enterprises in the manufacturing industry”. It uses the information available in the BR, and it may put on restrictions, for example that the enterprises included are active when the frame is constructed.

An annual survey collects data after the reference year, and a short-term survey collects data during the year (shortly after each month/quarter). If the frame is constructed shortly before sending out the questionnaires, that time is at the end of the year for the annual survey, and shortly before the reference year for the short-term survey. The latter may take further samples during the year. Anyhow, the frame errors are different for these two sets of statistics – unless the annual statistics deliberately use the same frame as the short-term statistics for the sake of agreement, compare Chapter 10.

The frame population is based on the information that is available at that time. For short-term statistics regarding year *t*, the SIC codes refer to year $(t-1)$ at best – more likely to year $(t-2)$ or possibly even earlier, depending on the production time of the statistics used and the frequency of updating. In the case of the manufacturing industry this normally depends on when PRODCOM information becomes available.

Note: PRODCOM is short for the French words "Production communautaire" meaning Community production.

5.3.3 Differences between the frame population and the target population

There are two types of differences between the frame and target population:

- differences for the population as a whole;
- differences within the population, affecting domains (sub-populations).

Another way of expressing this is the classification of for example an enterprise into surveys or into domains within a survey. (This could be manufacturing versus service industries, and industries within the manufacturing industry, respectively.) Those two cases will be dealt with in Sections 5.3.4 and 5.3.5, respectively.

A part of the target population may deliberately be left out of the survey, for example enterprises below a certain size may be cut off. The estimation for this part of the population has to be based on model assumptions, see Chapters 4 and 9. Administrative data may be useful, especially if there are variables strongly related to those of the statistics.

A different classification of frame "errors" is with respect to the time it takes until they are corrected. Some are simply due to time delays in the information from different sources. Such errors can be evaluated after updates. Other errors are either detected in special circumstances – like a survey or a change including that information – or (more or less) never detected. Those errors can hardly be studied; at the least they require special investigations. Small units especially may be subject to an error for a long time.

The updating procedure may sometimes be held back deliberately, as mentioned above in Section 5.3.2 for coherence between short-term and annual statistics in some Member States. Another example is for short-term statistics using the same set of classifications and size measures during the year, used in the UK in order not to add the effects of re-classifications to the within-year-changes. Both stratum and domain are ”frozen”, see further Sections 5.6.1 and 5.7.1.

5.3.4 Under- and over-coverage of the population

There are two types of deviations between the frame population and the target population:

- under-coverage: units belonging to the target population but not to the frame population
- over-coverage: units belonging to the frame population but not to the target population

There is an asymmetry between the two. A consequence of under-coverage is that observations are not collected for a part of the target population. This may imply a bias in the statistics. Over-coverage means that resources are used on uninteresting units. The over-coverage may be regarded as an "extra" domain of estimation, and one of the results (in comparison with no over-coverage) is an increase in uncertainty when estimating the "regular" domains. If the unit's membership of the target population is not checked, there may be a bias.

For both under- and over-coverage, the resulting inaccuracy depends on the amount of the coverage deficiencies, the ability to detect them, and the counter-actions taken in the estimation procedure.

Furthermore, there may be practical difficulties in distinguishing over-coverage and unit nonresponse. A unit outside the target population that receives a questionnaire may be more or less inclined to return it than a unit belonging to the target – it is easy to return, but on the other hand there seems to be no reason to fill in the questionnaire. Some questionnaires may be returned by the postal authorities because the address is no longer valid – that should, of course, be followed up. See Chapter 8.

5.3.5 Differences within the population

The reasoning that was used in the previous section for the whole population is to some extent also valid for each sub-population. However, under-coverage of one domain is over-coverage for another.

There are some different possibilities here for coverage deficiencies:

- remain undetected (for example an erroneous SIC code remains)
- detected for the sample (or more accurately for the responding units; for example the number of employees in the questionnaire)
- detected on the population level (for example a general update of SIC codes between sampling and estimation)

Again, the resulting inaccuracy depends on the amount of the coverage deficiencies, the ability to detect them, and the counter-actions taken in the estimation procedure.

5.3.6 Some comments on frame errors

Even if the construction of a frame population is easy in principle, there is much work in practice with the BR and the frame with regard to births, deaths, organisational changes, contradictory pieces of information, duplicates, mistakes, identification problems, time delays, etc. Identification is important, for example to eliminate duplicates due to different sources. Archer (1995) describes the maintenance of business registers, including some examples from New Zealand. One statement made is that identifying births typically involves a quarter of the total resources needed.

A close co-operation between the BR and the statistical surveys using it as a frame is important. This includes an understanding on both sides of the different uses. It also means a lot of work on single cases to handle them correctly both over time and in different surveys, for example in cases of reclassifications and reorganisations. Particular care is needed with large enterprise groups which have complex structures and span several different activities. Such entities may cut across different surveys, and the structures are subject to change. It is important that they are monitored closely so that changes can be picked up quickly and handled consistently. In the UK there is a Complex Business Unit to this end. A number of other countries have a similar organisation, some of them also being responsible for all survey data collection.

In the discussion of quality assurance for business surveys by Griffiths & Linacre (1995), frame creation, maintenance, and monitoring is an important part, including illustrations of births, deaths, and time lags.

The term *frame error* is not always a correct description – *coverage deficiency* is often more adequate, showing the consequence and not just blaming the frame, for example for not having included mergers in January 1998 in a frame constructed at the end of 1997.

5.3.7 Defining a Business Register covering a time period

The target population has reference times for the units that equal those of the variables, as mentioned above. This means, for enterprises and annual statistics for example, that the enterprises included should not be those that are active at the time of the frame construction but all enterprises that are active during the year, whether active the whole year or during a part of the year only.

If the frame is constructed at the end of the year (see discussions in Sections 5.3.2 and 5.6.1-5.6.2), the enterprises missing in the frame are "early deaths and late births", that is broadly those that are (i) no longer active according to the BR but have been active previously in the year, and (ii) not active in the BR but active later in the year. Moreover, with SIC codes referring to a different period than the target calendar year, there will be misclassifications.

This shows the frame deficiencies affecting statistics unless actions are taken. A special BR with the purpose of such actions is introduced below.

At some point after the calendar year it is possible – at least in principle and if the information needed has been kept – to combine information from the BR including time stamps, and possibly also from other sources, to derive a new Business Register that refers to the calendar year. In the case of enterprises, it includes all enterprises that have been active at some time during the calendar year. The values of the variables also refer to the full year. If the basic values have reference times that are points in time, some procedure is needed, perhaps a suitably chosen average of values before/during/after the year. The same is possible for a different period, like a quarter, but due to the time delay, such a register is less likely to be useful.

Sweden has some experience of a BR covering a calendar year and its use, illustrated in Sections 5.7.2-5.7.3. It is then regarded as the best knowledge attained. Statistics based on this BR and another, previous version are compared. This is one way to evaluate effects of frame errors. Furthermore, the improvement of the accuracy through using this BR should be considered together with the efforts involved, to see if the effort is cost-effective.

An “ordinary” BR shows the situation at some point in time, like a snapshot. However, considering that the rate of updating varies between variables and units, it is rather a mixture of snapshots of the units with regard to delineation, SIC code, size measures etc.

5.4 The target population: estimation and inaccuracy

5.4.1 Estimation procedures and information needed

As stated several times, the target population has reference times of basic variables like SIC code that are equal to those of the statistics. For example, both annual and short-term statistics referring to year *t* should be based on delineation of units and SIC codes of that year. The frame is based on a BR at a time too early to achieve this.

There are several possibilities at the estimation stage, with different ambitions for updating the information and, at the same time, with different results as to accuracy with respect to frame errors (coverage deficiencies). Whatever the procedure chosen, the resulting (in)accuracy needs to be measured.

A typical situation is a design with stratification by industry and size. A random sample is drawn for each stratum. The greatest size strata have the selection probability equal to one. The stratification into sets of SIC codes corresponds to the domains, each stratum being equal to (or more detailed than) a domain. Size is used in the stratification to improve accuracy.

The basic estimator of the total value of production, say, for a particular industry is then simply a sum over the size groups for that industry. The variance of the estimator is also computed by summing over these strata. The estimation procedure can use a Horvitz-Thompson estimator, expanding sample values by inverted probabilities of selection (in the case of full response), see further Chapter 2. This is so for the sampling unit and its domain as given by the frame. With a different observation unit, the contribution to a particular industry will also come from other strata, for example if enterprises are sampled and their kind-of-activity units are the observation units.

There are further possible estimators, depending on what information is available in addition to that in the frame. There are two main reasons to use further information:

- to reduce bias by including corrections and updates;
- to reduce variance through utilising auxiliary information.

The amount of further information may vary: it can be limited to the sample or it can be available for the population, for example in terms of further variables or an updated BR.

Some situations are described below in Sections 5.4.2-5.4.5. For estimation procedures, see Chapters 2-3 or the literature, for example for calibrated weights in generalised regression estimators see Deville & Särndal (1992).

5.4.2 Using the frame population only

The simplest estimation procedure is to keep to the frame population, that is, each unit keeps its domain of estimation as on the frame. As described above, each pair of point estimate and standard error is computed by summing over the corresponding strata.

This procedure can be used not only for classification but also for units that are in fact dead or otherwise not belonging to the target population, by treating them like nonresponse. If there is no renewal of the sample, such an estimation procedure can be regarded as including a model assumption on the relationship of under- and over-coverage: that they are equal in size. There is bias due to under- and over-coverage for the population as a whole and for each domain, unless the assumption is true. When the birth rate is high compared to the death rate, there is under-estimation and vice versa.

Care needs to be taken in using simplified assumptions. Investment provides a particular challenge. New units and ones which are growing are likely to be strong investors. Conversely units which are struggling and, as a result, diminishing in size will have little opportunity to buy new assets. Elvers (1993) discusses this for a survey based on a cut-off sample with the restriction 20 employees or more.

An alternative – leaving the frame information to some extent – is to identify the over-coverage and put variable values equal to zero for these units. If there is no renewal of the sample, there is then an imbalance, since over-coverage but not under-coverage is taken into account.

Illustrations: Table 5.3-Table 5.4 in Section 5.7.3 show an example of over- and under-coverage with a cut-off survey. The bias due to an old SIC-code is shown for an example in Figure 5.1 in Section 5.7.2.

5.4.3 Updating the sample only

If the units in the sample have their domain "checked" in the survey, interior movements and corrections can be taken into account by assigning each sample unit to its proper domain of estimation. This implies that the bias from this error source is eliminated. There is, however, an increased variance due to including this information – which may be a rare characteristic – based on sample information only. Chapter 3 provides formulas in its sections on domain estimation, for example a simple case in Section 3.1.2.

There may, in fact, be quite a difference in going from (i) the variance coming from a small set of "tailor-made" strata as indicated in Sections 5.4.1-5.4.2, to (ii) the variance derived from these strata and some further strata where a few units with actual values contribute to the variance together with a large number of nil values. This is a consequence of frame deficiency.

There are also exterior movements/corrections, units leaving and entering the population. An update in the first respect means for example identifying over-coverage and giving it a nil value. There is then an asymmetry if no action is taken for the under-coverage, as stated in Section 5.4.2. Either additional sampling or model assumptions are needed to estimate for units not in the population originally sampled, the frame population. A very simple model is to assume equal effects between over- and under-coverage, but this assumption is only likely to be realistic when the economy is stable – and not always even then.

For units included with probability one, changes can be made without affecting the variance, for example reorganisations can be taken into account and classification updates can be made, as long as each such unit represents itself only. However, care must be taken if surveys into different sectors are run independently. For example, if such a unit is reclassified from retailing to manufacturing, it could be removed from the retailing survey. A second action needs to be taken at the same time to ensure it is included in the manufacturing survey. There may be difficulties in doing this in practice.

Illustrations: The increase in variance (or rather its square root) when updating an old SIC-code based on sample information is shown for an example in Figure 5.1 in Section 5.7.2.

5.4.4 Utilising later BR information on the population

A situation with even more information is where there is a further variable for all units, not used in the design, or where there is renewed information on the original design variables.

One estimation method is so-called poststratification, where a stratification variable is added at the estimation stage. The calibration technique is an example of including such auxiliary information (possibly quantitative) to improve the estimation. This may lead to a reduction of both bias and variance. It is a model-assisted estimation method that is used for the surveyed part of the population.

Movements of units into the population are not included in the procedures just mentioned. They require model-based procedures with assumptions about these units. Again, there is an asymmetry to be overcome.

There are *illustrations* of changes in SIC code and number of employees from one year to the next in Table 5.2 and Figure 5.2, respectively. Table 5.1 has SIC code for a shorter period.

5.4.5 Utilising a BR covering the reference period

The technique of constructing a BR covering a period was described above in Section 5.3.7. The target population is here considered fully known. This is, of course, a simplification, since some errors will remain. This BR is, however, a considerable improvement over the version at the time of frame construction. From the estimation point of view, the situation with this BR covering the reference period is roughly the same as that in Section 5.4.4 in terms of methods and assumptions. This means for example that poststratification and calibration methods are available for interior movements.

Movements out of the population are identified, that is, the over-coverage is known. The under-coverage is also identified. The estimation has to be model-based for those units (unless there is time for further questionnaires), using for example similar units in the surveyed part of the population and/or administrative data. Again, the reasoning is based on this late BR covering a period showing the truth; in practice there are, of course, remaining deficiencies.

In Section 5.7.3, Table 5.3-Table 5.4 *illustrate* over- and under-coverage with a cut-off survey, and there is information on the "extra" units provided by the BR covering the calendar year.

5.4.6 Some comments on the BR and effects of coverage deficiencies

Discussions on the topic of quality of a BR are going on at the EU level (Eurostat 1998a). The connections between Business Registers and the statistics using them are getting stronger. There is an increasing interest in business demography, and regular work on quality assessment of business registers is taking place at some statistical offices. See also Struijs & Willeboordse (1995), Archer (1995), and Griffiths & Linacre (1995), already mentioned, and illustrations below.

The measurement of inaccuracy caused by coverage deficiencies may be undertaken in three different ways:

1) Review updating procedures of the BR to look at time delays. This will provide a broad indicator only, but it is available at the time when the frame is constructed.
2) Compare units on an updated BR with the BR used. Counts can be made of the number of units erroneously included or excluded. Likewise the number of units classified to the wrong domain of estimation can be evaluated.
3) Compute approximately the level of inaccuracy. Estimates can be made for the frame population and for the estimated target population, using a variable that is available at the population level (for example turnover from VAT, or salaries and wages or number of employees from PAYE). Whilst this method provides the most information it is the most demanding and resource intensive.

The illustrations in Section 5.5 are tied to the BR, and Sections 5.6-5.7 provide a range of illustrations for frames, although nearly restricted to the UK and Sweden. Most illustrations in Sections 5.6 and 5.7 belong to the first and second of the above methods. There are, however, a few examples on accuracy measures in Section 5.7 belonging to the third method. This is the preferable one, since a quality assessment should aim at the effects of frame errors (coverage deficiencies).

5.5 Illustrations – administrative data and business demography

Business Registers are dependent on administrative data and influenced by administrative rules, which may vary over time and, of course, between countries.

As an example, a birth in the BR can have different causes: there are pure births in the sense of new activity, and there are new registrations due to a new legal form or an enterprise reorganisation into several parts etc. According to a survey on different characteristics of new Swedish enterprises, about 54 % of the 1997 new BR enterprises were purely new, see SOS (1998); the figure for the previous year was 60 %. (These figures refer to enterprises with more than SEK 30 000 (approximately 3 500 ECU) in annual turnover, but the survey also covers smaller enterprises.) Statistics Finland (1996) gives similar results. The percentage depends on the BR system, of course, and it varies between countries and over time. Another way to study business demography is to utilise individual employment data together with the BR. A description for Sweden is given in Statistics Sweden (1995); the method is stated to be a transformation of original ideas from Denmark.

The dependence on administrative rules is illustrated in two tables. The first one shows the number of units in the Swedish BR by year, with some comments on considerable changes.

Year	Number of active legal units	Changes in Tax and VAT-rules in Sweden
1986	520 657	
1987	489 904	
1988	491 747	
1989	508 266	
1990	568 356	From 1990 includes units without activity code
1991	494 802	Change in VAT-rules
1992	493 690	
1993	493 070	
1994	553 290	New kind of tax (some influence on 1993 also)
1995	562 765	
1996	584 206	

The next table is a related one from the UK. The basis of the data collection by the ONS is the Inter-Departmental Business Register (IDBR), which was introduced in 1994 and became fully operational in 1995. The IDBR combines information on VAT traders and PAYE employers in a statistical register comprising 2 million enterprises, representing nearly 99% of economic activity. The register comprises companies, partnerships, sole proprietors, public authorities, central government departments, local authorities and non-profit making bodies. The main administrative sources for the IDBR are HM Customs and Excise, for VAT information (passed to the ONS under the Value Added Tax Act 1994) and Inland Revenue for PAYE information (transferred under the Finance Act 1969). Other information is added to the register if required for ONS statistical purposes. This table includes information only on VAT-based enterprises.

Notes: The counts of businesses below the VAT threshold representing voluntary registrations and with zero turnover are included in the two first parts of the table (1984-1993 and 1994-1995). Figures for the first part are counts of individual legal units. Counts for the second part show VAT-based enterprises consisting of one or more legal units. The third part (1995-1998) excludes units with zero VAT turnover and all enterprises without a VAT basis. The GBP is currently around 1.4 ecus.

Year	Number of legal units / enterprises	Percentage change in number	Change in VAT-registration date and threshold value in GBP	
1984	1 496 957		1984-03-14	18 700
1985	1 513 922	+ 1.1%	1985-03-20	19 500
1986	1 533 156	+ 1.3%	1986-03-19	20 500
1987	1 558 306	+ 1.6%	1987-03-18	21 300
1988	1 609 176	+ 3.3%	1988-03-16	22 100
1989	1 680 670	+ 4.4%	1989-03-15	23 600
1990	1 765 178	+ 5.0%	1990-03-21	25 400
1991	1 795 360	+ 1.7%	1991-03-20	35 000
1992	1 723 239	– 4.0%	1992-03-11	36 600
1993	1 671 611	– 3.0%	1993-03-17	37 600
			1993-12-01	45 000
1994	1 628 969	– 2.6%	1994-11-30	46 000
1995	1 606 067	– 1.4%	1995-11-29	47 000
1995	1 551 525		as above	
1996	1 537 645	– 0.9%	1996-11-27	48 000
1997	1 547 175	+ 0.6%	1997-12-01	49 000
1998	1 573 935	+ 1.7%	1998-04-01	50 000

5.6 Illustrations – time delays and taking frames

5.6.1 The UK Business Register

The UK register holds two classifications and two measures of size. A current value shows the latest position and is used to form the frame for the annual inquiries. A “frozen” value (updated only at the start of the year, before January selections, from the current values at that time) is taken through the year to ensure consistency throughout the year for sub-annual inquiries. Thus the annual frame relates to a later period than the short-term frame, the UK concentrating on accuracy for structural statistics in preference to congruence with short-term surveys.

The register is updated from a number of sources during the year:

i. PAYE Updates. Tapes are received from the tax authority every quarter giving details of new units, closures and changes of structure.
ii. VAT Updates. A weekly tape is received from HM Customs and Excise containing details of births (new registrations), deaths (deregistrations) and amendments. Enterprises with no local units or PAYE units have an employment imputed from the VAT unit turnover using the turnover per head figure appropriate to the classification.
iii. Survey Information. Size and classification data update only the current classification.
iv. Visits by the Complex Business Unit (see Section 5.3.6). These are supplemented by desk profiling within the Business Register Unit.

The births, deaths, and restructurings picked up from these sources are actioned immediately. Classification and size amendments affect only current values unless a unless a business is in the process of being profiled or a significant error is found.

Updating of the register takes place through the year from quarterly sources such as PRODCOM, but the main update is in August from the Annual Employment Survey (to be incorporated into the annual structural survey from 1998). The results of the update will drive selection for the sub-annual inquiries for the following year.

The sources of information used to update the IDBR are listed by variable below:

I. Turnover. The VAT administrative system is the main source. Survey data are used from the distribution ("trade") and services sectors but rarely from elsewhere. Enterprises with no VAT or survey information have a turnover value imputed from employment information.

II. Employment. The preferred source is the Annual Employment Survey (to be incorporated into a new annual structural survey from 1998). Employment information comes from the PAYE ("Pay-as-you-earn") tax administrative system if Annual Employment Survey data are not available. Enterprises with no employment information (either from PAYE or from AES) have employment imputed from turnover.

III. Classification information comes from a variety of sources. The following priority applies:

   A. Complex Business Unit
   B. PRODCOM/Retail Inquiry/Financial inquiries
   C. Annual Register Inquiry
   D. Short Period Turnover Inquiry
   E. Other business surveys
   F. Builder's Address file
   G. VAT
   H. PAYE

   The annual register inquiry is a new survey which will replace so-called "register proving" from 1999. The Builder's Address File contains information on construction businesses from the Department of the Environment, Transport and the Regions' (DETR) construction industry surveys.

Care must be taken when using two administrative sources such as PAYE and VAT to update the BR to ensure that erroneous information is not taken on and used in producing estimates. When a new PAYE unit is identified with 20 or more employees, an attempt is made to match it with a VAT unit or a local unit elsewhere on the register. If no corresponding unit is found, the unit is sent a register proving form and excluded from all estimates until its validity is confirmed. Likewise a new VAT unit would be matched with PAYE, and proving undertaken if no corresponding unit can be found. Extensive matching is carried out for units with fewer than 20 employees, but there is no proving for these units due to resource and compliance constraints. Small unmatched PAYE units in VAT exempt industries and corporate PAYE units are added to the register without proving.

The annual structural survey samples are drawn at the end of October each year. The short-term surveys are drawn dynamically each month or quarter. Samples from the short-term inquiries are drawn from the frozen field whilst the annual inquiries select from the current fields. Samples are stratified by industry and size. The measure is usually employment. The size groups for the annual structural sample for the production industries and for the Monthly Production Inquiry are shown below.

Annual Production	Monthly Production
0-9	0-9
10-19	10-49
20-49	
50-99	50-149
100-249	150+
250+	

5.6.2 The Swedish Business Register

This description refers to the middle of the 1990's, mainly before the EU Regulations came into Swedish use (Sweden became a Member State in 1995). The Swedish BR obtains information on births and deaths from the National Tax Board every second week. The number of employees is updated through several sources. The two main ones are the Tax Payroll and a special questionnaire to multiple-location enterprises, both once a year. There is also information from the surveys of Statistics Sweden. For Divisions 10-37 of the Swedish SIC 1992 that is harmonised with NACE Rev.1 at the four digit level, there is an annual survey roughly at the local unit level that is an important source for the SIC code (using output information, including data on commodities).

There is a modified version of the BR, called the Statistical Register (SR), which is used as the frame for business surveys. Some units on the SR consist of a set of legal units. They are the smallest ones for which balance sheet and profit and loss data can be obtained. They are essential to the Financial Accounts Survey, and they are included in other frames for coherence. There are about 60 such large statistical units, consisting of more than 400 legal units. In the following, the term enterprise will be used to mean such units whenever they occur and legal units otherwise. (This enterprise definition is somewhat different from the EU one. An enterprise includes more legal units in some cases, and fewer in other cases; there should be further enterprises with several legal units. The number of such enterprises has, however, increased recently.)

In the sampling system, most samples are drawn in November (and some in May). The SR can then be expected to describe the situation at the end of September as to active enterprises and local units. The number of employees refers to the spring this year, t, for multiple-location enterprises (BR questionnaires) and to December last year, year $(t-1)$, for single-location enterprises (PAYE information). Single-location enterprises born in year t normally have 0 employees in the BR that year. Hence surveys that require a minimum of for example 10 or 20 employees do not cover births in year t.

The samples obtained are used for that year by annual surveys and for the next year by short-term surveys (some sampling is also made in May). All surveys use industry (the SIC code) for stratification. Most surveys also stratify by size, and the size measure is mostly the number of employees. The size groups in the surveys here are six, with the two top ones totally enumerated (that is 200 employees or more). They are based on enterprises in Divisions 10-37 with at least 10 employees. They also include (roughly) local units with at least 10 employees in Divisions 10-37 belonging to enterprises in other Divisions for functional statistics, but that part of the population is disregarded here for simplicity.

number of employees:	10-19	20-49	50-99	100-199	200-499	500+

5.6.3 Some comparisons between UK and Sweden

UK and Sweden have similar routines in several respects, for example in using both PAYE and VAT as sources, and by putting extra emphasis on the BR quality around October with regard to frames. Samples for annual surveys depend on that frame, and so largely do short-period samples. Stratification by industry and size is used.

There are also differences, for example UK uses dynamic sampling for short-period inquiries and Sweden runs surveys with a cut-off limit. There are differences between units, for example the enterprise concept and the extent of applying the kind-of-activity units. UK has a special team for complex businesses, and Sweden has a special BR covering a calendar year.

5.7 Illustrations – changes between frames and their effects

5.7.1 Differences between UK current and frozen classifications

The matrix in Table 5.1 shows for the UK how enterprises are classified on the BR in relation to current and frozen SIC classification. It reveals the extent to which the frozen classification is wrong at one point in time (September 1998, following the take-on of the 1997 Annual Employment Survey (AES) information). It should be remembered that short-term inquiries select from the frozen field for purposes of consistency during the year. The matrix should be interpreted in the following way:

Rows: the figure at the end of the row shows the percentage of businesses that have remained in the division of their frozen classification following the AES update (and any other information (for example from PRODCOM) received during the year). It also shows the extent to which businesses will be reclassified out from an industry.

Columns: the figure at the bottom of the column shows the percentage of businesses currently classified to a certain division which were classified to that division in the frozen field also. It also shows the extent to which businesses will be reclassified in to an industry.

The matrix reveals a relatively small amount of reclassification in terms of numbers of businesses with reclassifications in or out of less than 3 % for nearly all industries. It would be interesting to see the analysis carried out on employment too. (Note: It would also be more interesting to have a full year matrix, but this is not possible for 1997 or earlier.)

The industry with the highest percentage of inward reclassifications stored up is division 31. Here, 95.2 % of the enterprises in the current field are also in the frozen field, so 4.8 % (308 businesses) will be added when the current field is copied over into the frozen field. The industries which provide the most enterprises are divisions 32 and 33. Conversely, 2.1 % of

	Current Sic92																																
Frozen Sic92	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	40	41	45	Total	% on diag-onal
10	231	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	232	99.6
11	0	363	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	5	371	97.8
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0
13	0	0	0	60	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	65	92.3
14	1	0	0	2	1349	1	0	0	0	0	0	0	0	1	0	0	3	0	1	1	0	0	0	0	0	1	0	0	0	0	7	1367	98.7
15	0	0	0	0	1	8676	0	0	0	0	1	0	0	0	6	0	0	0	1	5	0	0	0	1	0	0	4	0	0	0	1	8696	99.8
16	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	100.0
17	0	0	0	0	1	1	0	6874	32	8	1	2	6	0	4	4	1	0	1	2	0	1	0	0	1	0	10	1	0	0	5	6955	98.8
18	0	0	0	0	0	0	0	32	8427	33	0	0	1	0	0	0	1	0	1	1	0	1	0	1	0	0	9	1	0	0	2	8510	99.0
19	0	0	0	0	0	0	0	1	12	1354	0	1	2	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	1	1375	98.5
20	0	0	0	0	0	0	0	2	1	0	9042	4	2	0	1	9	7	0	23	3	0	1	0	0	1	4	47	0	0	0	78	9225	98.0
21	0	0	0	0	0	0	0	2	0	2	3	3130	37	0	2	10	1	0	1	2	1	1	0	0	0	0	3	0	0	0	1	3196	97.9
22	0	0	0	0	1	0	0	12	1	0	1	88	31428	0	2	2	1	0	3	6	2	4	1	1	0	2	7	0	0	0	5	31567	99.6
23	0	2	0	0	0	0	0	0	0	0	0	0	0	295	2	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	301	98.0
24	1	0	0	0	1	3	0	3	0	2	1	0	4	0	4532	19	4	0	11	5	2	2	0	4	0	0	6	0	0	0	5	4605	98.4
25	0	0	0	0	0	1	0	3	2	1	9	5	10	0	13	7345	10	0	26	22	0	7	1	4	5	5	14	2	0	0	17	7502	97.9
26	1	0	0	0	9	0	0	5	1	0	1	1	2	0	9	33	5969	2	10	5	0	4	1	2	3	0	8	0	0	0	42	6108	97.7
27	0	1	0	0	0	0	0	0	0	1	3	2	0	1	3	3	0	2920	109	9	0	5	0	0	1	0	9	2	0	0	11	3080	94.8
28	0	0	0	0	1	0	0	3	1	1	8	2	9	0	4	37	6	56	30686	163	1	24	3	9	11	13	47	2	0	0	84	31171	98.4
29	0	0	0	0	1	2	0	3	0	3	2	1	4	0	3	18	3	9	138	15961	2	22	8	27	12	9	14	0	0	0	49	16291	98.0
30	0	0	0	0	0	0	0	0	0	0	0	0	6	0	2	1	1	0	3	4	1923	29	13	50	0	0	3	0	0	0	1	2036	94.4
31	0	0	0	0	0	0	0	1	0	0	0	1	4	0	2	4	2	3	12	19	6	6154	18	22	4	0	3	0	0	0	34	6289	97.9
32	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	3	1	0	5	5	5	49	3362	12	3	0	4	0	0	0	13	3464	97.1
33	0	0	0	0	0	1	0	0	0	1	1	1	2	0	5	9	0	0	13	36	40	69	33	6340	2	2	5	0	0	0	6	6566	96.6
34	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	4	1	1	14	7	1	6	0	0	3487	2	3	1	0	0	4	3533	98.7
35	0	0	0	0	0	0	0	0	0	0	2	0	1	0	1	2	1	2	9	11	1	2	2	2	7	3481	2	0	0	0	6	3532	98.6
36	0	1	0	0	1	5	0	24	8	9	59	10	14	0	9	51	18	5	63	24	5	17	7	21	5	6	20888	2	0	0	62	21314	98.0
37	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	1074	0	0	1	1081	99.4
40	0	2	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	256	0	30	293	87.4
41	0	0	0	0	0	8	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	91	3	106	85.8
45	1	5	0	0	9	8	0	10	1	1	179	1	7	0	5	86	61	5	268	139	5	63	17	28	4	14	76	2	1	0	209343	210339	99.5
Total	235	374	0	62	1378	8707	29	6975	8486	1416	9313	3250	31543	298	4608	7640	6093	3004	31399	16435	1994	6462	3467	6527	3546	3539	21165	1088	258	91	209817		
% on diago-nal	98.3	97.1	0.0	96.8	97.9	99.6	100	98.6	99.3	95.6	97.1	96.3	99.6	99.0	98.4	96.1	98.0	97.2	97.7	97.1	96.4	95.2	97.0	97.1	98.3	98.4	98.7	98.7	99.2	100	0.0		

Table 5.1 Comparison of frozen and current SIC-codes in 1998 on two digit level

the enterprises with the frozen classification in division 31 will be leaving the industry. Divisions 32 and 33 feature again, now as the main destination industries .

The industries with the largest amount (in percentage terms) of outward reclassifications awaited are divisions 40 and 41; 12.6% and 14.2% of businesses respectively will be leaving the industry. Due to the fact that there are not that many businesses operating in these industries this does not represent many businesses (37 and 15 respectively).

5.7.2 Differences within the Swedish population one year apart

There are four main data sources which can be used to study changes among enterprises in Divisions 10-37 with at least 10 employees. First, the BR covering a calendar year, here 1995. Second, the frame for the short-term survey, both in November 1994 and in November 1995. The frame for this survey is essentially the same as that of the annual survey, but used one year earlier. Third, the register where observations and imputations from the annual survey 1995 have been added for comparative purposes. Fourth, there is administrative data, PAYE and VAT.

The main files used in this Section are the frames from November 1994 and 1995, and they include enterprises with at least 10 employees in Divisions 10-37. Hence, differences between situations one year apart are shown. They correspond to the frames for the short-term and annual statistics for 1995. The short-term statistics largely keep their classifications, but the annual statistics make new ones, so the differences in industry in the statistics will be based on data two years apart. The files are at the enterprise level.

The SIC code has five digits, the fifth being a Swedish addition, which is rarely different from zero. Changes are for convenience studied by using all five digits, without regard to letters, making differences based on the first two digits a bit unequal. Table 5.2 shows by row to which digit the SIC codes agree for enterprises in 1994 and 1995. The column shows size group in 1995. Nearly 500 units have a change in SIC code. There are not considerable differences between size groups as to percentages of changes.

		Size group for 1995 (number of employees)						
		10-19	20-49	50-99	100-199	200-499	500+	Total
Number of equal digits in SIC code	0	40	36	19	2	4	1	102
		2.11	1.81	2.32	0.49	1.39	0.52	
	1	37	42	19	8	1	1	108
		1.95	2.11	2.32	1.96	0.35	0.52	
	2	64	50	19	9	5	4	151
		3.38	2.52	2.32	2.21	1.74	2.09	
	3	26	22	8	4	4	1	65
		1.37	1.11	0.98	0.98	1.39	0.52	
	4	24	24	11	4	0	5	68
		1.27	1.21	1.34	0.98	0.00	2.62	
	5	1704	1813	742	381	273	179	5092
		89.92	91.24	90.71	93.38	95.12	93.72	
Total		1895	1987	818	408	287	191	5586

Table 5.2 Comparison of SIC-codes 1994 and 1995 with regard to size 1995. In each cell, the upper figure shows the frequency, and the lower figure shows the column percent.

The changes here are considered fairly “normal”. (There is an exception for the division 22, with considerable changes with the new SIC code. If changes from 1995 to 1996 had been chosen to overcome the SIC code effect, there would have been a greater influence from collecting commodity data in a new nomenclature and in a new way.)

Changes for large enterprises will have a considerable effect on institutional (enterprise-based) statistics. The effect for short-term functional statistics *may* be fairly small: if there are two kind-of-activity units with roughly the same size, the change in primary activity of the enterprise may be caused by small changes in relative size between the two kind-of-activity units.

The effects on the two-digit-level of institutional statistics are shown in terms of absolute numbers on the vertical axis of Figure 5.1, using the newer number of employees (from the 1995 frame). There are 27 domains of estimation. Six of these domains are unaffected. Two of them are affected by more than 5%. A more detailed level is, of course, more sensitive.

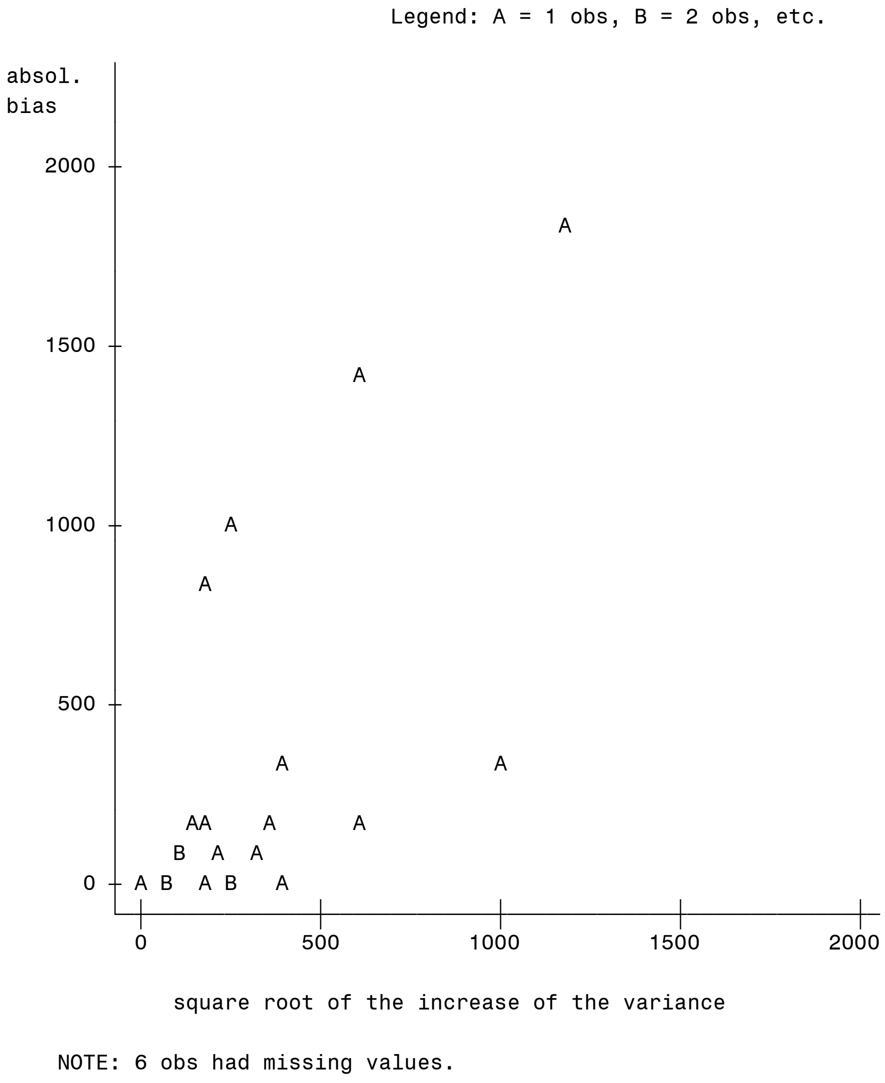
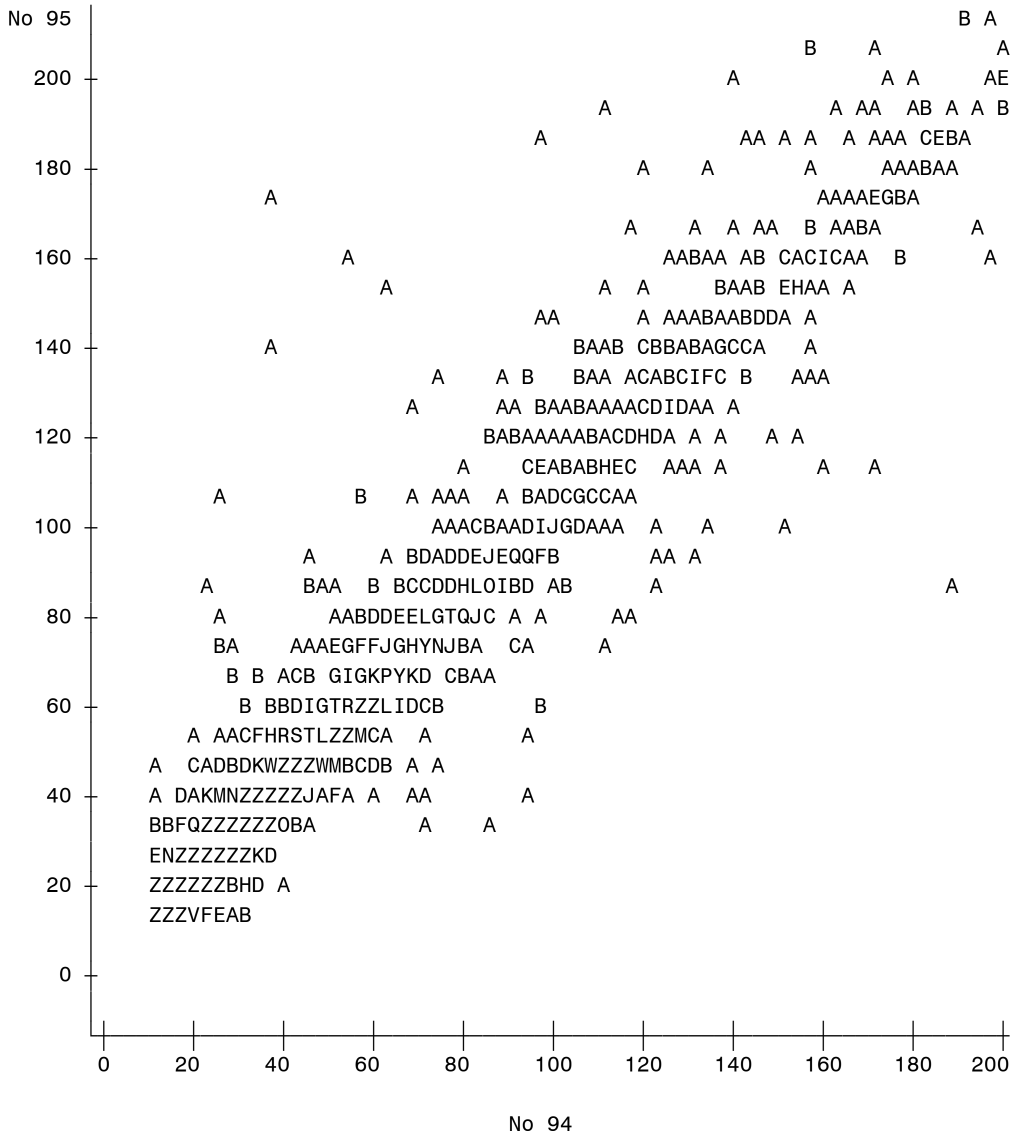


Figure 5.1 Comparison of absolute bias due to the old SIC code in the frame and the square root of the increase of the variance when updating the sample only.

The horizontal axis of Figure 5.1 shows the increase in the square root of the variance in using the SIC code of the sample instead of the SIC code of the frame. The sample size has been derived by a simple Neyman allocation in the 1994 frame with the precision criterion 1% for the number of employees. The details are left out, as the aim is just a simple illustration.

The line "$y = x$" in Figure 5.1 corresponds to the two mean squared errors being equal. Points above that line (8 in number) correspond to industries that get a smaller mean squared error if the bias is eliminated by updating the SIC code for the sample. For points below the line (13 in number) the increase in variance when there are contributions not only from the "tailor-made" strata is so large that the resulting mean squared error is higher than the original one.

Legend: A = 1 obs, B = 2 obs, etc.



```
NOTE: 2844 obs hidden.  471 obs out of range.
```

Figure 5.2 Number of employees in 1995 versus that in 1994, according to frames.

Now to size and change in size, first illustrated in a simple plot, Figure 5.2. The figure is restricted to the number of employees being at most equal to 200, that is it shows the sampled part of the population. A greater spread of a survey variable within strata can be expected when the stratification is based on the size of the 1994 frame than would have been the case with the size of the 1995 frame. The plot indicates that some units will have much higher or much lower values than the stratification indicates. The accuracy of the estimates will be lower than they would have been with a more up-to-date size.

A related illustration is a cross-classification of the size groups in the two frames. The resulting table (not included here) shows that somewhat more than 10% of the enterprises move upwards or downwards by one or possibly two size classes. The years 1994 and 1995 were such that the movements upwards dominated over those downwards. To remain in the same size class is, of course, by far the most frequent case, seen in somewhat less than 90% of the enterprises.

5.7.3 Differences for the population as a whole; Sweden

The data files used to study differences for the population as a whole are those mentioned in the previous section. The BR covering 1995 is in this context considered as the final result. The number of employees is a convenient measure of the effects. There are two disadvantages, however: there is no contribution from enterprises with no employees and there is a "full" contribution from enterprises which were active for only part of the year.

A count of the number of employees in small enterprises (less than 10 employees) in the BR covering 1995 shows that 7.4 % of the total number of employees is there, making 55 thousand employees below the cut-off. There are 679 thousand employees above the cut-off. According to the two frames, the numbers are 639 and 657 thousand employees, respectively, the differences being due to both differences in units and differences in reference times.

The over- and under-coverage of each of the two frames are shown in Table 5.3 on the first row (bold italics) and the first column (bold), respectively, in terms of number of enterprises. The group 'below' includes both small and non-active enterprises. It should be noted that the figures given are mainly the result of a "blind" match-merging. Enterprises that belong to the totally enumerated group on one occasion and the below group on another are likely to have gone through some re-organisation, taken into account by the survey.

Looking at the annual survey, where the BR covering a calendar year is used to produce the statistics, the percentages of additions relative to the 1995 frame are of the same overall order. There are some differences in procedures. The under-coverage found in that survey is checked to avoid double counting. On the other hand, enterprises may falsely be dropped at an early stage as over-coverage and then "return" as under-coverage. A set of enterprises of ancillary character is "picked up" from the Financial Accounts Survey.

Out of the 567 enterprises that were in the BR covering 1995 but not in the 1995 frame, 2 were well above the cut-off but in other industries, 209 were not active, and 358 were below the cut-off (62 of these without employees).

Consider now the sampled part only, but in more detail. First in Table 5.4, over-coverage is

groups in the	groups in the frame 1994			groups in the frame 1995		
BR cov. 1995	below	sampled	tot.enum.	below	sampled	tot.enum.
below	-	***522***	***13***	-	***176***	***0***
sampled	**1 276**	5 251	6	**550**	5 980	3
tot.enum.	**20**	37	461	**17**	11	490

Table 5.3 Over- and under-coverage of the frames 1994 and 1995

shown with number of units and the number of employees in thousands as measured by the frame and by the BR covering 1995. Then the under-coverage is shown with number of units, number of employees according to the BR covering 1995, and in relative terms summarised for three variables: number of employees, salaries and wages, and turnover from VAT. The figures here refer to the whole of Divisions 10-37. The relative effect on an industry level may, of course, be different, larger or smaller.

	over-coverage		under-coverage		
	units	empl. 1000's	units	empl. 1000's	three variables
1994 frame	522 ent.	10 → 2	1 276 ent.	→ 25	around 3.0 to 3.7 %
1995 frame	176 ent.	1.9 → 1.5	550 ent.	→ 12	around 0.8 to 1.7 %

Table 5.4 Over- and under-coverage of the sampled part, frames 1994 and 1995.

5.8 A few summarising conclusions

The BR and the frame derived from it provide a fundamental basis to the statistics. The frame population should be defined with regard to the target population, and the units of the BR should correspond to the statistical units. This is in line with the EU regulations.

Obviously, correct delineation and classification of units are important for the domains of estimation. Size information is often used to improve accuracy; deficiencies in size information will make the estimation procedure less efficient and cause troubles with outliers etc. The distinction between frame errors and other non-sampling errors is not always clear-cut as measurement errors may be related to unclearly or erroneously specified units, and nonresponse and over-coverage are not always easy to distinguish.

It is not only the frame – the BR at the time when the frame is constructed – which is important, but also to what extent the estimation procedure takes later information into account. This is so both for units that represent only themselves and units that represent others as well. It is normally the case that

- the inclusion of new information for the sampled part of the population implies a higher variance in comparison with the ideal situation with a perfect frame, but
- to disregard the information normally implies a bias.

When assessing the quality of the statistics, the resulting accuracy is the main aim. Time delays for new units and updates are indicators, but indicators only (Section 5.4.6).

6 Measurement errors

Chris Skinner, University of Southampton

6.1 Nature of measurement error

6.1.1 True values

Measurement error is defined relative to the value of a given variable (that is a question) reported by a given respondent. The basic assumption is that there exists a *true value* of this variable for this unit, so that there is no ambiguity in the definition of the variable. Given this assumption, the *measurement error* is defined as the difference between the reported value and the true value. This is not an operational definition, of course. Even if it is accepted that there can be no ambiguity in the definition of the true value, there may be no operational way for an agency to obtain the true value with certainty. Instead, various indirect methods may be used to detect measurement errors as described in this chapter.

6.1.2 Sources of measurement error

In this report measurement errors will be equated with 'response errors', that is errors arising because the respondent fails for some reason to provide the true value desired. Errors on the part of the data collection agency, for example falsely transcribing values from questionnaires or misrecording values reported by telephone, will be treated as processing errors (see chapter 7). Errors in auxiliary variables recorded on a business register will, furthermore, be treated as frame errors (see chapter 5). These errors may be attributable simply to out-of-date information on register variables but may also arise for similar reasons to response errors, that is because a business fails for some reason to provide the true value of the variable required.

Response errors may arise from three sources.

True value unknown or difficult to obtain

Sometimes the business may keep information according to different definitions, for example many businesses maintain accounts according to different financial years and it may be difficult to report values with respect to a different time period, for example a calendar year, requested by the agency. In such circumstances the business may report the value of the variable according to the closest definition available, for example the business's financial year.

Sometimes the business may not keep the information required, for example both the 'value' and 'quantity' of gas or electricity purchased, as asked in ONS's Annual Business Inquiry. Alternatively, the business may be unwilling to go to the effort required to retrieve the information. In such cases the value may be guessed or the question left blank. The occurrence of such measurement errors may therefore be indicated by high rates of item nonresponse on a question.

Such errors may have a particular effect on 'other' categories. For example, the ONS's ABI requires that expenditures in different areas should sum to the total expenditure reported. One of the last expenditure questions is for ‘other services purchased’. It is possible that this is used as a ‘balancing box’, according to which businesses simply work out what expenditure for the year has not already been accounted for.

Misunderstanding of question or other slips

Instructions on questionnaires may be misunderstood or simply not read. A common example of an error is the reporting of a value in the wrong units. For example, a question may ask for a value to be reported in units of thousands of pounds. A true value of £2,488,500 should therefore be reported as 2,489. A business may, however, erroneously report the figure as 2,488,500. Some forms include boxes within which digits should be recorded for scanning and businesses may complete these wrongly, for example writing ‘NIL’ through the boxes. The questions themselves may also be fundamentally misunderstood. For example, a construction firm might record the value of ‘retail turnover’ on the ABI as the firm’s expenditure on construction of retail outlets, whereas the true value should be zero.

Errors in information used by the respondent

Finally, it is possible that the information used by the respondent, for example from a business information system, is itself subject to error.

6.1.3 Types and models of measurement error

Four kinds of measurement error may be distinguished.

Continuous variables: major occasional errors

Examples of major occasional errors are the occasional reporting of values in the wrong units (for example in single currency units rather than 1000 currency units) or the occasional recording of expenditure under the wrong heading (so that expenditure under one heading is greatly reduced and expenditure under another heading is greatly increased). These errors will often be identifiable under close inspection as outliers (Lee, 1995). These are outliers which arise from error rather than outliers which are unusual but correct. If possible they should be detected and treated as part of the editing process (see section 6.3.3).

A stochastic model for such error in a measured variable Y would be that Y equals the true value with probability $1-\varepsilon$ and is drawn from a very different distribution with probability ε, where ε is a small number, for example 0.01.

Continuous variables: misreporting of zeros

A specific instance of major error is the misreporting of zeros. One example is the setting above where expenditure is recorded under the wrong heading so that expenditure under the correct heading may be erroneously zero whereas expenditure under another heading may be erroneously non-zero. Such errors may cancel out under aggregation of headings.

Other erroneous reportings of zero may arise when information is unavailable or difficult to obtain, a question is left blank and then imputed as zero. In this case, measurement error is closely related to item nonresponse (see Case Study 1 in Section 6.3.1).

Continuous variables: other error

Guessing of values and errors due to minor differences in reference periods might be expected not to lead to major errors but rather to errors which might be represented by the 'classical error model'

$$Y = y + e \tag{6.1}$$

where Y is the reported value, y the true value and e is the measurement error drawn from a continuous probability distribution. Sometimes the distribution of the errors might reasonably be supposed to be centred about zero, for example under honest guessing by an experienced reporter, so that the measurement error may be viewed as approximately unbiased. Sometimes, bias may be expected.

Categorical variables: misclassification

Measurement error in categorical variables involves misclassification. The basic model in this case involves a misclassification matrix with elements q_{ij}, the probability of classifying category i as category j. The diagonal elements of this matrix should be close to one and the off-diagonal elements small.

6.2 The contribution of measurement error to total survey error

6.2.1 Total survey error

Let Y_k be the reported value for the k^{th} sample unit and let y_k be the corresponding true value, assumed to be well-defined. Then $Y_k - y_k$ is the measurement error for sample unit k and the contribution of measurement error for all sample units to a weighted estimate $\sum_s w_k Y_k$ is given by $\sum_s w_k (Y_k - y_k)$. This contribution to total survey error reflects not only measurement error but also processing, coding and imputation errors.

In order to assess the magnitude of the contribution of $\sum_s w_k (Y_k - y_k)$ to total survey error (see Section 1.2.1), it is necessary to conceptualise the distribution of this term and to estimate the characteristics of this distribution. The distribution of $Y_k - y_k$ usually involves the specification of a measurement error model as in (6.1). The measurement error distribution in such models might be conceived of in terms of hypothetical repeated measurements (Groves, 1989, p.15). For example, a respondent might provide different guessed values if asked (hypothetically) the same question repeatedly, or different individuals might complete a form differently under (hypothetical) repeated mailings to a firm. The distribution might also be conceived of in terms of the distribution of errors across businesses. For example, an error arising because a respondent refers to the business's financial year rather than a calendar year may not change under repeated questioning, but it may be possible to interpret the distribution of errors e in the model in (6.1) as reflecting the distribution of financial years (in their impact on the survey variable) across businesses.

Given a measurement error model, the distribution of the total survey error can be conceived of as reflecting the joint distribution arising from measurement error, sampling and nonresponse. If E denotes expectation with respect to the joint distribution, the bias and variance arising from measurement error (and associated processing, coding and imputation errors) may be expressed as

$$\text{Bias} = \mathrm{E}\left(\sum\nolimits_s w_k (Y_k - y_k)\right) \tag{6.2}$$

$$\text{Variance} = \mathrm{E}\left(\sum\nolimits_s w_k (Y_k - y_k)\right)^2 \tag{6.3}$$

The assessment of these is considered in Section 6.4 below. For the purpose of quality measurement, the primary interest will be in total survey error and an overall measure of quality is

$$\text{Mean squared total survey error} = \mathrm{E}\left[\sum\nolimits_s w_k Y_k - \sum\nolimits_P Y_k\right]^2 .$$

6.2.2 Bias

The bias in (6.2) may arise from all kinds of measurement error. For example, a systematic tendency to underreport certain miscellaneous costs may lead to downward bias in the estimation of total miscellaneous costs. A tendency to report according to an earlier financial year rather than a requested calendar year may lead to downward bias for variables which exhibit upward trends over the time period concerned.

6.2.3 Variance inflation

The variance inflating impact of measurement error is likely to be most important for the largest businesses in the completely enumerated strata. Such businesses do not contribute at all to the sampling variance, but random errors in their reported values may have a significant impact on the total variance of the survey estimates. This is considered further in Section 6.4.3.

6.2.4 Distortion of estimates by gross errors

Usually, it is assumed that the total survey error and its components are normally distributed so that the distribution can be summarised by bias and variance. An exception may arise with gross errors which are not detected or treated. Gross errors for individual businesses may seriously distort estimates, especially estimates for domains based on small numbers of observations, one (or more) of which is subject to gross error.

6.3 Detecting measurement error

6.3.1 Comparison at aggregate level with external data sources

Survey estimates may be compared with aggregate figures from another source, such as another survey, an administrative source or trade organisation data. Such a comparison may reveal bias from measurement error, although it may be difficult to disentangle measurement error bias from nonresponse bias and it may be difficult to determine to what extent the difference between estimates is attributable to error in the survey of interest or error in the other data source.

Case Study 1. Comparison of mail survey with interview survey

In the 1980s Statistics Sweden conducted an annual survey on forestry (logging) among private owners (as opposed to large corporations, the government or the Church). The private owners make up about 50% of all forestry in Sweden. This survey was done by a conventional mail questionnaire design and involved a sample of 7,000 such owners (owning less than 1,000 hectares each). The aim was to estimate at the national level, among other quantities, the total volumes (in million cubic meters) logged by final felling (that is a whole area is cut down), thinning (selected trees only) and miscellaneous felling (in ditches, under power lines etc). Because of concerns about quality, it was decided in 1988 to divide the survey into two parts on an experimental basis: a mail questionnaire was distributed to about 4,500 owners while about 2,500 owners were included in an interview survey, about 100 local forestry experts performing the interviews. The results are given in the following table.

	π-weighted estimate of proportion of owners doing activity		Estimated volume (million cubic meters)	
	Mail	Interview	Mail	Interview
Final felling:	20%	21%	17.6	19.0
Thinning:	32%	39%	9.7	11.3
Miscellaneous:	18%	38%	1.9	3.7
Total logging:			29.2	34.0

The estimated volume for the mail survey tends to be less than for the interview survey, especially for the miscellaneous category. This may be explained by the much greater numbers of zeros (owners not undertaking the activity) in the mail survey, especially for the miscellaneous category. Many of these zeros represent either measurement error (the failure to report actual activity) or item nonresponse (a blank return where an actual return may be difficult). Final felling is easy to identify and quantify (for example lots of paperwork is involved to get a permit), while thinning and particularly miscellaneous logging are harder to identify, quantify and remember. It was concluded that the quality of the results from the mail questionnaire was unacceptable and the survey was changed to an interviewer mode from 1989.

6.3.2 Comparison at unit level with external data sources

A more useful comparison is possible if the respondent records can be matched to records from another source such as a tax register, containing related variables. Such comparisons might only be made with a subset of sample records, for example the responses of just the businesses in the completely enumerated stratum might be compared with information in publicly available annual reports. Gross errors might be detected in values which do not follow the normal relationship with variables in the external source. Differences in definitions between the two sources, in particular differences in reference periods, will often complicate such comparisons, however. It may also be that the external source, for example an audited set of company accounts, only becomes available after the survey estimates have been published, so that measurement error estimates can only be made retrospectively.

Case Study 2. Comparison of questionnaire responses with values on VAT register

The survey on 'domestic trade in the service sector' at Statistics Sweden aims to estimate quarterly turnover by industry (4 digit NACE) in the service sector. A probability sample of legal units is drawn from the Business Register (BR) and a questionnaire is mailed to these units.

In 1997 a study was made to find out whether the mail questionnaire could be replaced by data taken directly from the VAT register. Such a shift would reduce costs considerably, for Statistics Sweden as well as for respondents, and at the same time make it possible to shift from a sample of about 4,500 to a total enumeration of about 110,000 enterprises.

Two estimates of turnover by 4-digit NACE were compared. The first was a π-weighted estimate from the original survey observations. The second was a modified estimate, with the questionnaire observations replaced by the corresponding VAT observations (except in the take-all strata).

Differences between the estimates were reasonably small in most NACE groups compared with the random variation in the survey. However, in some NACE groups the differences were much larger than one would expect from the random variation. For 114 legal units the π-weighted difference between questionnaire and VAT data exceeded 50 million SEK. About one third (37) of these were selected for a telephone interview to find out the reasons for the discrepancies. For practical reasons the interviews had to be done during the holiday season in the summer, and only 21 interviews were completed. Nevertheless, a lot was learned from these interviews:

1) In 10 cases (legal units) the large discrepancies were due to the choice of unit. These legal units turned out to be part of multi-legal unit enterprises. The turnover in the sample cases may be reported to the VAT register from another legal unit within the same multi-unit enterprise, and this VAT-reporting unit may even be an out-of-scope unit, for instance a manufacturing unit. In some cases the selected unit reported zero turnover while the corresponding VAT turnover was substantial. In some cases it was agreed (with the respondent) that the questionnaire turnover was indeed the correct one while the VAT turned out to be the correct figure in other cases.
2) In 3 cases the respondents had by mistake given the wrong numbers (turnover) on the questionnaire. This had been corrected during the discussion, making questionnaire- and VAT data coincide.

3) Two cases were due to data entry errors made by Statistics Sweden but not detected by editing.
4) Two cases were due to errors in NACE classification in the BR. The respondents had reported ‘Manufacturing’ instead of the service sector code found in the BR. These units had been classified as over-coverage in the survey and given value of turnover equal to zero.
5) One case, a wholesale trade agent (NACE = 51.1) had included as turnover the whole traded turnover instead of only its own turnover as requested in the questionnaire.
6) Two cases were traced to misunderstanding of the questionnaire.
7) One case was due to reference period problems. This enterprise was involved in a 6 month long project. The VAT payments were divided into six monthly equal sums while actual payment took place on one or two occasions. It so happened that the ‘questionnaire-turnover’ was attributed to another quarter than the one in the study while the VAT data seemed to be very consistent from month to month.

It is clear that such comparisons with external sources can reveal many sources of error in addition to measurement error. In particular the most striking additional type of error in this study consists of frame errors arising from problems in delineating units. Such comparisons may also suggest methods for improving quality. This study suggests, for example, that VAT data may be useful for editing. A large difference between questionnaire responses and VAT turnover would be a good reason for a telephone contact.

6.3.3 Internal comparison and editing

A simpler approach is to examine the internal consistency of the values reported in the survey as part of the usual editing process (Hidiroglou & Berthelot, 1986; Pierzchala 1990; Granquist and Kovar, 1997). Thus, one may check accounting identities, for example where components sum to a total, and inequalities, for example that some variables are positive. Comparisons may be made with values reported in previous surveys by the same respondent. For example, a variable with month to month variation normally not in excess of 5%, which suddenly changes by 1000% is a likely case of gross measurement error. See chapter 7, on processing errors, for further discussion.

6.3.4 Follow-up

When edit constraints are failed, there are generally two options. First, the reported values may be modified so that they do obey the constraints, for example following the procedure of Fellegi and Holt (1976). Second, the respondent may be followed up in order to clarify the reason for the failed edit constraint and hence to establish, if necessary, a value with reduced measurement error. Such follow-up may be expected to provide more information about the nature and size of the measurement error. It may be selective, that is only values considered likely to have a non-negligible effect on the statistical estimates might be followed up.

Follow-up can range from a simple telephone call to check a single value through to a more detailed reinterview, aimed at establishing the sources of information used as well as the respondents' understanding of questions and instructions. Dippo, Chun & Sander (1995, p.295) refer to this as a *response analysis survey*. Such a survey may reveal measurement errors directly, for example through misunderstandings displayed, or may suggest subgroups for which the quality of the data may be worst. For example, respondents might be asked whether their responses were based on memory or involved reference to appropriate information sources. The proportion of respondents using memory might be taken as an indicator of poor data quality and might be compared between different subgroups of businesses.

Reinterviews appear to be relatively uncommon in European business surveys. An illustration of response variability is provided by a study of Friberg (1992) in which reinterviews arose by accident! He reports on a Statistics Sweden survey on environmental investments and costs in Sweden. A reminder was distributed at some point to those enterprises that had not yet responded. Five enterprises among those receiving the reminder had in fact sent in their questionnaires just one of two days before. It so happened in those five cases that a different person at the enterprise than the one who had already responded (and then possibly gone on holiday - this happened in the summer) filled in the questionnaire. This made it possible for Statistics Sweden to compare the two versions from each of the five enterprises. Very large differences were found between the responses of the pairs of respondents from each of the five enterprises. This seems to reflect the large degree of error in measuring a variable such as environmental investment, which is difficult to define and quantify.

6.3.5 Embedded experiments and observational data

Randomised experimental designs may, in principle, be used to detect measurement error bias by comparing alternative measuring instruments (Biemer & Fecso, 1995, p.268). For example, different form designs or different modes (for example mail versus telephone) might be assigned randomly between different respondents. See Case Study 1 in Section 6.3.1 for an example.

Randomised assignment may often be difficult to implement in practice. For example, although an agency may request that a form be answered by a particular category of staff, it may be difficult in practice to enforce this. It might therefore be difficult to implement a randomised experiment comparing the effect of using, for example, management versus clerical staff as respondents. It may, however, be possible to record observational data on the category of staff responding in an ongoing survey. The fact that the allocation of staff is not experimentally assigned makes the interpretation of differences in the survey outcomes between different categories of staff more difficult, because of potential confounding with other variables, but not impossible (Biemer & Fecso, 1995, p. 269).

6.4 Quality measurement

6.4.1 Quality indicators

There are several ways that problems in the quality of responses to a particular question may be revealed:

a) high rates of failure of different edit constraints involving the variable;

b) high rates of item nonresponse may indicate difficulties in answering the question and potential measurement error;

c) unexplained large variation between survey occasions;

d) spontaneous reports on difficulties from respondents;

e) a response analysis survey (Dippo *et al*., 1995) may reveal misunderstandings or the frequent use of memory in answering a question;

f) subject matter understanding of the nature of the question, for example investments are harder to quantify than the number of people employed.

There are also several indicators for problems with the whole questionnaire:

a) response burden in terms of time and effort;

b) number of people involved in responding to the survey;

c) change of person responsible for filling in the questionnaire;

d) proportion of late/delayed responses.

Quality indicators derived from such sources may be useful for monitoring quality and for comparing quality between questions and between surveys. They may suggest possible directions of bias but are unlikely to provide much help in the assessment of the magnitude of the bias or variance of total survey error.

6.4.2 Assessing the bias impact of measurement error

Where specific sources of measurement error are concerned, bias may be assessed by modelling the mechanism leading to error. For example, the effect of businesses using their own financial year rather than the requested calendar year might be adjusted for by applying a trend model within industrial categories to the sample businesses which do not use the calendar year. Or the impact of businesses allocating activity to erroneous headings might be assessed by estimating the probability of misclassification between headings.

Sometimes it may be possible to conduct experiments (see section 6.3.5) to assess the bias impact of alternative measuring instruments, for example different form designs or mail surveys versus telephone surveys. Differences between measuring instruments only reflect different biases, however, and do not necessarily provide accurate estimates of absolute biases.

Another approach to bias assessment is through comparison with external sources (see section 6.3.1). Again, it may not necessarily be possible to decide which source is least biased and, moreover, measurement error biases will generally be confounded with other sources of bias, such as nonresponse.

The ideal way to assess bias is to conduct reinterviews with the sample to establish the true values. Such an exercise faces, of course, many practical obstacles (Biemer & Fecso, 1995, p.270).

6.4.3 Assessing the variance impact of measurement error

Variance estimators designed to estimate the sampling variance (see chapter 2) may also be expected to capture an important component of the variance of total survey error attributable to measurement error. Consider, for example, the classical error model in (6.1), where the reported value Y is determined from the true value y by $Y = y + e$ where e is the measurement error. Consider a single stratum *h*, within which the true values and errors are independently distributed with common variances σ_{yh}^2 and σ_{eh}^2 respectively. Let $\overline{Y}_h$ be the mean of the measured variable *Y* for the n_h sample units in the stratum and let μ_h be the mean of the true values *y* for N_h population units in the stratum. Thus $\overline{Y}_h$ is the survey estimator of μ_h. Assuming simple random sampling within the stratum, the variance of the total survey error $\overline{Y}_h - \mu_h$ across both the sampling and measurement error distributions is obtained as

$$\mathrm{var}(\overline{Y}_h - \mu_h) = \sigma_{yh}^2 \left({}^{1}\!/_{n_h} - {}^{1}\!/_{N_h} \right) + \sigma_{eh}^2 / n_h \,,$$

The usual estimator $v_s(\overline{Y}_h)$ of the sampling variance of $\overline{Y}_h$ is the sample variance of *Y* within the stratum multiplied by $\left(1/n_h - 1/N_h\right)$ and this has expectation

$$\begin{aligned} \mathrm{E}\left[v_s(\overline{Y}_h)\right] &= (\sigma_{yh}^2 + \sigma_{eh}^2)\left({}^{1}\!/_{n_h} - {}^{1}\!/_{N_h} \right) \\ &= \mathrm{var}(\overline{Y}_h - \mu_h) - \sigma_{eh}^2 / N_h \end{aligned}$$

The estimator is therefore biased downwards, failing to capture the component σ_{eh}^2 / N_h arising from measurement error, but the bias will be small if N_h is large. A conservative approach is to remove finite population corrections from the variance estimator (that is replace $\left(1/n_h - 1/N_h\right)$ by $1/n_h$). This is likely to be too conservative, however, especially for completely enumerated strata. To obtain an improved variance estimator it is necessary to estimate the variance σ_{eh}^2 of the measurement error. This might be attempted via a reinterview survey (Biemer & Fecso, 1995, p.265). If not, some kind of sensitivity analysis is likely to be necessary.

The contribution σ_{eh}^2 / n_h of the measurement error to the variance above assumes independent measurement errors. If measurement errors for different businesses are positively correlated then this will tend to inflate the variance. It is important therefore that the variance estimator is based on reporting units between which independent reporting is a reasonable assumption. If, for example, a single respondent provides responses for several enterprises, the measurement error could be correlated between these responses and so the set of enterprises should be treated as a single reporting unit for the purpose of variance estimation.

7 Processing errors

Pam Davies, Office for National Statistics

7.1 Introduction to processing error

Once survey data have been collected from respondents, they pass through a range of processes before the final estimates are produced. These post-collection operations can have effects on the quality of the survey estimates. Errors introduced at this stage are called processing error. Processing error can be divided into two categories: systems error and data handling error.

The topic of processing error is just one component of non-sampling error. Non-sampling errors, including processing errors, affect not only data from sample surveys but also administrative and census data. Processing errors, along with other nonsampling errors, may lead to biases and increases in the variance.

This chapter concentrates on describing the various components of processing error in the context of business surveys. Some suggestions are made for reducing the effect of processing errors on data quality. The report is illustrated by examples, from the UK and Sweden, of research to measure and minimise processing error.

7.2 Systems error

Systems errors are errors in the specification or implementation of systems used to carry out surveys and process results. One source of systems error is automated data capture. Systems errors typically affect either all or particular classes of estimates.

The impact of systems errors on data quality is influenced by when the errors are discovered. The impact of the errors on data quality needs to be evaluated and compared with the cost of correcting the error, both in terms of human resources and a possible delay in the release of the data, before making a decision whether to correct the error.

Systems errors which are discovered before the beginning of data collection are more easily corrected than errors which are identified in the course of the survey. With the use of computer assisted data collection, sometimes program errors are not detected until after data collection has started.

Systems errors later in data processing sometimes are not detected until later on, or, at worst, until after results are published, leading to the need to publish corrections. Clearly such errors are potentially very serious.

7.2.1 Measuring systems error

In order to measure the effect of a systems error, the parts of the system that are incorrect need to be corrected. The estimates need to be produced on both the incorrect and correct systems, and the difference in the results from the two systems needs to be compared.

7.2.2 Systems error: two examples

7.2.2.1 Sampling in the ONS

From about 1991 to 1994, probability proportional to size (PPS) sampling was used in some business surveys run by the UK Central Statistical Office (CSO), notably in the quarterly capital expenditure and quarterly stockbuilding inquiries. The system of sample selection was implemented on CSO's business register system, and was specified so that a random number was generated, and each business was represented by a part of the random number range proportional to its size, and was selected if the generated number fell into its interval. The coding in the program did not, however, follow this procedure exactly, and in 1994 it was discovered that the selection probabilities were not as intended. The suggested solution was to work out what selection probabilities were implied by the selection procedure, and to use those to produce an unbiassed (but possibly rather variable) set of survey estimates.

The result of this episode was a general distrust of PPS sampling for business surveys, and, although a corrected selection algorithm is available, the method has been mothballed in ONS (the successor to CSO) since then.

7.2.2.2 Variable formats in computer programs

When a computer program is being written, variables may be allocated certain fixed formats, and say for a particular variable the format is defined to be an integer with two digits. At the time a value above 99 is considered impossible. In time, values above 99 become possible and occur, but nobody amends the format. The system chops values to store them within the stated format, and does so without warning, for example 123 simply becomes 23. The statistics then do not move as expected. After a while somebody realises the cause!

7.2.3 Minimising systems error

Systems errors are minimised by the use of quality assurance and testing procedures as the system is written. Where appropriate, the use of harmonised methods across surveys enables the same well-developed and tested program code to be used for processing data in all the surveys. This reduces the scope for programmer error by reducing the amount of code to be written, and frees up resources for developing and testing other parts of the system.

7.3 Data handling errors

Potential sources of data handling errors range from processes used to capture and clean the data to techniques used for the final production of estimates and the analysis of the data. The main sources of data handling errors are:

- data transmission: this covers errors arising in the transmission of information from the field, where data are collected, to the office where the data are subjected to further processing;
- data capture: 'the phase of the survey where information recorded on a questionnaire is converted to a format which can be interpreted by a computer';

- coding: ‘the process of classifying open-ended responses into predetermined categories’ (Kasprzyk & Kalton, 1998);
- data editing: ‘a procedure for identifying errors created through data collection or data entry using established edit rules’ (Kasprzyk & Kalton, 1998). Data editing also refers to the automatic correction of certain errors where the error is (apparently) identifiable or where the cost of checking it manually exceeds the benefit over automatic correction;
- any process that is applied to the data, from the identification of outliers to the seasonal adjustment procedure, can introduce processing error. This processing error is not caused by the method itself, but by the incorrect application of the process.

This report discusses errors introduced at the data transmission, capture, coding and editing phases of the survey.

7.4 Data transmission

For most business surveys, data transmission from the field is via postal questionnaire. In this case, transmission errors are unlikely to cause a significant problem because the data should arrive intact. In some instances, data may be faxed or given over the telephone and in these cases the scope for error increases. Faxed information may be illegible, and information given over the telephone may be misunderstood, or recorded wrongly by the survey workers. In both these cases, if there is any doubt, the recorded value should be checked with the respondent before it is captured.

A relatively new development, at least for ONS business surveys, is the use of touch-tone, rather than mailing, for data transmission. Clearly there is scope for respondents to either fail to operate the system correctly, or to press an incorrect button. To minimise the risk of errors, the system should be designed so that respondents are required to confirm their return. Gross errors are detected in the editing phase, but smaller errors may otherwise pass undetected.

7.5 Data capture

A variety of methods may be used to ‘capture’ data. These include:

- keying responses from pencil and paper questionnaires;
- using scanning to capture images followed by automated data recognition to translate those images into data records;
- keying by interviewers of responses during computer assisted interviews;

and these are discussed in turn below.

7.5.1 Data keying from pencil and paper questionnaires

The traditional method of data capture for business surveys is the keying of responses from pencil and paper questionnaires onto computer by a centrally located data entry team. This is a very labour intensive task, which has now been replaced on many surveys by more modern technologies. Some modes of data collection such as computer assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI) enter the data onto computer in the course of the interview.

Data keying is used on many postal surveys where pencil and paper questionnaires are the simplest way to collect information. ONS is investigating the potential of other methods of data capture including scanning and automated data recognition to reduce the number of surveys where data are captured in this way.

7.5.1.1 Measuring error occurring during data keying

The accuracy of data keying can be measured by either comparing a batch of entered data with the original questionnaires or more commonly by re-entering the batch and comparing the two sets of data. Lyberg & Kasprzyk (1997) give a range of examples with error rates varying from 0.1% to 1.6%. Any new methods of data capture must have error rates at least as good as these to maintain the quality of survey data.

7.5.1.2 Minimising error occurring during data keying

Methods of minimising errors during data keying include:

- checking regular batches of questionnaires for keying errors;
- in-built edits in computer assisted transmission can identify keying errors;
- checking all data entry work of new staff until they reach an acceptable level of accuracy.

7.5.2 Data capture using scanning and automated data recognition

The potential cost savings offered by the use of scanning and automated data recognition over traditional data keying has led to increasing interest in this technology. In ONS scanning is being used for some business surveys. For example, the last Census of Employment carried out in the UK used scanning equipment to capture all the data resulting in quicker processing and a lower cost for a very large survey. Other organisations who have investigated the use of scanning and automated data recognition for data capture include Statistics Sweden (Blom & Friberg, 1995), and Statistics Canada (Vezina, 1996).

The stages in the data capture process are:

- **Scanning**

The questionnaires are separated into single sheets and fed into the scanner which stores the image of each page as a TIF file. The preparation of questionnaires for the scanner can be fairly labour intensive (Elder & McAleese 1996) since any staples need to be removed and the questionnaire correctly aligned. The storage of images of questionnaires has the additional advantages of providing rapid access to questionnaires if any queries arise and reducing the need for storage of large volumes of paper questionnaires.

- **Form Out**

In many data recognition systems the image of the original printed questionnaire is removed electronically from the image of the data filled in by the respondent. This reduces the computer memory needed to store the image of the data and clarifies the image for automated data recognition.

- **Automated data recognition**

Different methods are used to extract the data from the image depending on the type of information being captured. These include:

- Bar code recognition (BCR). Used to read bar codes, for example serial numbers on paper questionnaires. Very accurate.
- Optical Mark Recognition (OMR). Used to read responses in tick boxes. Over 99% of items are (presumably correctly) recognised by the system.
- Optical Character Recognition (OCR). Used to read machine-printed text. Over 99% of items are (presumably correctly) recognised by the system.
- Intelligent Character Recognition (ICR). Used to read hand-written characters. For hand written numerical information 65%-90% of question responses were recognised. This figure is lower for hand written text information; as a result ICR is rarely used for collecting such information.

The recognition figures quoted above are from Statistics Sweden's experience of automated data recognition as reported in Blom & Friberg (1995). It must be emphasised that technology is developing quickly in this area so that the accuracy of automated data recognition systems can be expected to improve.

7.5.2.1 Measuring error associated with scanning and automated data recognition

Automated data recognition may introduce errors into data when characters are incorrectly recognised; for example the numbers 3 and 8 may be confused, as may the numbers 1 and 7. If the system is more likely to confuse a 3 for an 8 than vice versa, and similarly for the numbers 1 and 7, then these errors could cause an upward bias in the survey estimates. Some of these errors may be detected at the editing stage but some inaccuracies may slip through.

The accuracy of automated data recognition may be compared with keyed data entry by processing a batch of forms in both ways and comparing the resulting data. Elder & McAleese (1996) report the results of such a comparison where they found that for some questionnaires the accuracy achieved by the automated recognition system was at least as high as that achieved by the keyed data entry process.

7.5.2.2 Minimising error associated with scanning and automated data recognition

The most effective way to ensure high quality data capture using automated data recognition is to design forms that are easily scanned and interpreted by the data recognition process. Vezina (1996) provides a useful discussion of aspects of form design that influence data quality. These include:

- the characteristics of the paper – it needs to feed easily into the scanner;
- the colour of the ink – scanners pick up some colours better than others and this can be used to enhance the images of the data;
- page identifiers;
- registration points – marks on the form which enable the system to align the scanned image with what it's expecting;

- definition of zones of data to be captured – this is particularly important for parts of the form where the respondent is asked to write in numbers or letters. The provision of boxes encourages the respondent to print characters in capitals that are easier for the system to recognise than manuscript;

and to these we could add one which Vezina does not mention:

- instructions asking the contributor to provide data in the required format.

7.6 Coding error

The aim of coding is to transform open-ended, textual information into categories that can be used in data analysis. In the business survey field, the commonly used coding classification is NACE Rev. 1, but in the UK this is replaced with the comparable Standard Industrial Classification 1992 (CSO, 1992).

A major use of coding in business surveys is on the business register. In the UK, businesses provide a description of their activity, which needs to be coded according to the Standard Industrial Classification. In some business surveys open-ended descriptions, for example of commodities, are required that need to be coded according to a product classification.

The accuracy of coding is heavily dependent on the skills of coders, so there is the potential for introducing both bias and variance during the coding process.

Coding has two stages:

- the development of a classification or coding frame. This coding frame is known as a nomenclature or dictionary and is accompanied by a set of coding instructions. Nomenclatures need to be frequently revised so that they represent the full range of possible categories;
- written or verbal responses to survey questions are coded into categories. This coding may be:
    - ➢ strictly manual where the human coder looks up the codes in the dictionary;
    - ➢ computer assisted where responses are available in electronic form or typed into a computer and some purpose-written software suggests a range of possible codes. The human coder either selects one of these codes or edits the verbal description and asks the computer to suggest further possible codes;
    - ➢ completely automated. In completely automated coding the survey responses are available in electronic form or entered into a computer and the computer software allocates the code.

7.6.1 Measuring coding error

The impact of different coders on data quality can be assessed in terms of consistency (or reliability) and accuracy compared to a standard.

7.6.1.1 Consistency

A consistent coding system will give the same code for items in the same category. Computer automated systems are by definition completely consistent since given the same description of a category they will allocate the same code.

Different human coders implement coding rules differently, whether consciously or subconsciously, so they may allocate different codes to the same job description.

The consistency of coding systems can be measured by asking a set of different coders to code a common list of job descriptions and calculating the proportion of all paired comparisons of codes where the coders agree (Kalton & Stowell, 1979).

7.6.1.2 Accuracy

Although automated systems are completely consistent they have another less desirable feature: they may not allocate the best code to a description, that is, the code may not be an accurate one. Automated coding systems rely on the matching of text strings; if the matching is not exact then the assignment of codes may not be accurate. The accuracy of codes can be measured by comparing codes allocated by standard coders with those allocated by an expert coder, who is presumed to be infallible.

7.6.1.3 The impact of coder error on the variance of survey estimates

In manual coding and computer assisted coding different coders may allocate different codes to the same description. In particular each individual coder may unconsciously over-allocate businesses to some codes and under-allocate them to others. This is known as correlated coder error. The errors in the codes allocated by a particular coder may lead to bias in the estimate of the proportion of businesses in a given industry group for industries coded by that coder. However since for many surveys coding is shared over a number of coders, if the errors made by coders are different the impact of these individual biases on the final survey estimates may cancel out. In this case although the final survey estimates may not be biased the variance of the estimates will be increased. The overall bias is reduced as the number of different coders increases, so in some surveys the code list is provided with or as part of the questionnaire, so that each respondent codes their own answer. This minimises correlated coder error at the expense of a potential increase in measurement error (see chapter 6).

7.6.1.4 The risk of coder error introducing bias in survey estimates

Bias will be introduced into survey estimates if at least some coders systematically assign incorrect codes to certain occupations. One scenario where this may occur is in computer assisted coding where the computer suggests a preferred code which the coder may accept or reject. If there is a tendency for coders to accept the suggested code even when it is incorrect then the coding error may introduce bias into the survey estimates (Bushnell 1996).

7.6.2 Minimising coding error

The impact of coder error on data quality can be minimised by:

- the effective training of coders in using the coding system;

- well designed, up-to-date coding systems;
- in manual and computer-assisted coding systems, coders need to be supervised and the quality of their coding checked regularly. In some cases coders may be unsure which code to allocate and these queries will need to be referred to supervisors and in some cases researchers for reconciliation;
- some surveys (or more localised experiments) code information more than once using different coders and compare the resulting classifications to help resolve cases where there is some doubt as to the true code.

Useful references on coder error include Lyberg & Kasprzyk (1997).

7.7 Data editing

Granquist (1984) described editing as having three goals:

- to provide information about data quality;
- to provide information to help bring about future improvements in the survey process; and
- to clean up possibly erroneous data.

Checks used to identify suspicious data items are called edit rules. These include:

- range or validity checks – is the data item in the valid range for the data?
- consistency checks – is the data item consistent with other data provided by the respondent either in that interview/questionnaire or on a previous occasion ?
- routing checks – has the respondent answered the correct questions? This forms a large part of editing checks for pencil and paper questionnaires.

Computer programs are used to implement these edit rules either on-line during the data entry process (integrated editing) or in a batch process which produces a list of suspect data items for manual review.

Suspicious data items can be classified into fatal edits or query edits. Fatal edits identify clearly erroneous data whereas query edits identify data that are implausible.

In addition to different types of edit rules there is a variety of different approaches to editing:

- editing can compare different items of data for a given individual (is this item consistent with the other items?) or compare the same item for different individuals (is this data item much higher than the others?);
- editing can be conducted on aggregates (do the summary statistics or estimates for this batch of data look suspicious?) or on individual data. Suspicious batches of data can then be subdivided and the aggregate editing process repeated until the error(s) are narrowed down to individual data ;
- editing can be manual, by inspection of paper forms before or during data entry, or automated.

For general discussions on editing see Granquist (1995), Lyberg & Kasprzyk (1997), and Granquist & Kovar (1997).

7.7.1 Measuring the impact of editing on data quality

Different organisations, and indeed individuals within organisations, have different editing policies. There is consensus on the importance of correcting fatal errors where data are clearly erroneous. However some argue that surveys, particularly business surveys, are over-edited, and that much of the editing conducted to resolve query edits has little impact on the quality of estimates and therefore should be reduced. This would have a large impact on the cost of running surveys: editing can absorb as much as 20-40% of total survey budgets (Granquist & Kovar 1997). If the resources devoted to editing were reduced this would free staff to concentrate on minimising other sources of survey error which might have a greater potential impact on data quality.

Others argue that since it is impossible to pre-specify all the uses to which data will be put, the potential impact of inconsistencies in the data on estimates cannot be assessed. Data should therefore be edited until they are internally consistent, particularly if one output of a survey is a data set to be stored at an external archive that may be used by secondary analysts.

7.7.2 Minimising errors introduced by editing

Editing can introduce bias into survey estimates if it is based on pre-conceived ideas of what the data ought to look like which turn out in practice to be untrue. Editing may also artificially reduce the variance of survey estimates if real extreme values are incorrectly adjusted towards the mean of the distribution. This can result in over-optimistic claims about the precision of survey estimates.

Strategies to minimise errors introduced by editing include:

- involving subject matter specialists in the editing process so that edits are appropriate for the data;
- using standardised editing code for questions that are used on a range of surveys;
- testing program code used in editing by examining what happens to businesses with particular combinations of data values;
- feeding back information about data quality to the survey, questionnaire and edit design stages so that possible amendments to questionnaires, field procedures and edit rules that would improve data quality can be discussed.

7.8 An example of error at the publication stage

Production of many different official statistics, and in particular monthly statistics, is often subject to tight time constraints. All stages of the production process are then carried out with no time to spare. One of the steps to be taken quickly is moving a table into the press release. In comparison with the previous table a new month is added, and previous months may be revised.

In Sweden recently, a new month was added to a table in a press release and the revision for the previous month was overlooked. Several earlier months were shown in bold as revisions. Hence, the earlier figure for the previous month may be read as confirmed, and it is less accurate than it should be. The lesson is that the less manual typing of figures the better; tables should be moved as a whole, or an automatic procedure for generating them from the final data should be used.

8 Nonresponse errors

Chris Skinner, University of Southampton

8.1 Introduction

Nonresponse arises when a sampled unit fails to provide complete responses to all questions asked in the survey. Errors arising from nonresponse may be considered as an extension of errors arising from voluntary sampling, as discussed in Section 4.2, since the failure to volunteer information may be viewed as a form of nonresponse. Nonresponse errors are treated here as distinct from frame errors, as discussed in Chapter 5. In particular, sampled units which fail to respond but are outside the target population (ineligible) are treated as frame errors. In addition, noncoverage (that is units in the target population but outside the sampled survey population) is treated as a frame error.

8.2 Types of nonresponse

8.2.1 Patterns of missing data

Unit nonresponse arises when a unit fails to provide any data for a given round of a survey. There are two broad reasons for such nonresponse:

(i) *noncontact* – the form may not reach an appropriate respondent for various reasons, for example change of address, failure of the postal system, failure to forward from within the business;

(ii) *refusal* – the form does reach an appropriate respondent but the respondent does not return the form.

Unit nonresponders may be classified into two types according to the information available about the unit to the agency:

units which have never previously responded (these will consist primarily of smaller units which are sampled afresh at each survey occasion, or those newly recruited to the sample in rotating schemes) – for such units the only information available may be that recorded on the frame;

units which have previously responded (*wave nonresponse*) – these units will usually consist either of completely enumerated units which are sampled on every occasion or else larger units which are sampled over several occasions in a rotation design – patterns of nonresponse over the rounds of the survey might be denoted XXOXOOXX, for example, where X denotes response and O nonresponse and the most recent round of the survey is on the right.

Item nonresponse arises when a form is returned from the unit but responses to some questions are missing. Such missing data may arise, for example, because questions were overlooked or because the information required to answer the question was not available to the respondent. A particular problem in business surveys is the separation of item nonresponse from zeros. Respondents will often leave blank answers to questions about amounts, for example the value of production in a certain category, when the answer is zero.

8.2.2 Missing data mechanisms

In order to assess the errors which may arise from nonresponse it is necessary to establish a statistical framework within which the mechanism of nonresponse may be considered. Formally, nonresponse may be represented by 0-1 *response indicator variables* of the form

$$R = \begin{cases} 1 & \text{if value is recorded (response)} \\ 0 & \text{if value is missing (nonresponse)} \end{cases}$$

Unit nonresponse may be represented by a series of indicator variables R_k, defined for each unit k in the sample. This definition may be extended in various ways. To allow for repeated rounds of a survey, one may define variables R_{tk} for occasions t and units k. Item nonresponse may be represented by a series of response indicators, one for each variable for which missing values may occur. There is a number of alternative statistical frameworks within which the nonresponse mechanism may be represented. See Lessler & Kalsbeek (1992, Chapter 7) for a literature review.

The *deterministic approach* assumes that response indicator variables R_k are defined for all units k in the population and that their values are fixed. Thus, in the case of unit nonresponse, it is supposed that the population is divided into two 'strata': the respondents who always respond and the nonrespondents who never respond. The nature of the errors arising from nonresponse will depend on how well the estimation methods used to handle nonresponse compensate for differences between these two strata.

The *stochastic approach* treats the response indicator variables R_k as outcomes of random variables. A number of different stochastic frameworks is possible. In the case of unit nonresponse, one approach is to treat the set of respondents (those sample units for which $R_k = 1$) as a random subsample of the selected sample obtained through a process analogous to two-phase sampling (Särndal & Swensson, 1987). The nature of errors arising from nonresponse then depends on assumptions about how the subsampling occurs.

In the remainder of this report a stochastic approach is adopted, corresponding to modern statistical modelling. Both the response indicators R_k and the survey variables y_k are conceived of as outcomes of random variables and assumptions about the missing data mechanism are represented through assumptions about the joint distribution of the R_k and the y_k. This approach is particularly flexible for handling different kinds of nonresponse, for example both unit nonresponse and item nonresponse, and for extending to an integrated framework which allows for both nonresponse and measurement errors.

The above framework is very general and in order to make useful progress in assessing nonresponse errors or in adjusting for nonresponse it is necessary to make more specific assumptions about the nature of the missing data mechanisms. Three terms will be useful for describing such mechanisms.

Missingness is said to occur *completely at random* if R_k is stochastically independent of the relevant survey variables. For example, if unit nonresponse in a survey of production is being considered, this condition would imply that businesses with low levels of production would be as likely to respond as businesses with high levels of production. This condition is a very strong one and may arise only rarely in practice.

Missingness is said to occur *at random given an auxiliary variable* (or variables) x_k if R_k is conditionally independent of relevant survey variables given the values of x_k. Suppose, for example, that x_k is a measure of size, such as employment or turnover, available on the frame. In a survey of production, nonresponse would occur at random given the size variable if nonresponse is unrelated to production amongst firms of any given size. The distribution of nonresponse could vary, however, between firms of different sizes. This assumption is generally less stringent than the assumption that data are missing completely at random. It is also an assumption which underlies many adjustment methods by judicious choice of measured auxiliary variables.

A missing data mechanism which does not occur at random given available auxiliary variables is said to be *informative* or *non-ignorable* in relation to the relevant survey variables. Consider, for example, item nonresponse on a complex variable, for which the higher the value of the variable, the more work will tend to be required of a business of a given size to retrieve the information. In such circumstances, it may be that even after controlling for measurable factors, such as size of the business, the rate of item nonresponse tends to increase as the value of the variable increases. Item nonresponse on this variable would therefore be informative in relation to this variable.

8.3 Problems caused by nonresponse

8.3.1 A basic setting

The problems caused by nonresponse will clearly depend on the way nonresponse is treated. For convenience of exposition, a simple business survey setting is considered where stratified simple random sampling is employed and where, in the absence of nonresponse, the population total t of a survey variable y is estimated by the expansion estimator

$$\hat{t} = \sum_{h=1}^{H} N_h \ \bar{y}_h .$$

Here, $\bar{y}_h$ is the sample mean in stratum h, N_h is the number of businesses on the frame in stratum h and H is the number of strata. Perhaps the simplest way of treating both unit nonresponse and item nonresponse is to employ the same estimator with $\bar{y}_h$ replaced by the mean across all responding units in stratum h which provide responses to this variable. The latter mean is denoted $\bar{y}_{rh}$, where the subscript r indicates that this estimator is based upon respondents data. The estimator of the total is then

$$\hat{t}_r = \sum_{h=1}^{H} N_h \ \bar{y}_{rh} .$$

8.3.2 Bias

Within the setting in Section 8.3.1, the expectation of $\hat{t}_r$ may be expressed as

$$\mathrm{E}(\hat{t}_r) = \sum_{h=1}^{H} N_h \mu_{h,R=1}$$

where $\mu_{h,R=1}$ is the mean of the survey variable in stratum h amongst those who respond (R=1), and this expression may be compared with the expectation of $\hat{t}$ in the absence of nonresponse.

$$\mathrm{E}(\hat{t}) = \sum_{h=1}^{H} N_h \mu_h$$

where μ_h is the mean of the survey variable in stratum h. The difference between these two expectations determines the bias arising from nonresponse.

$$\text{bias}(\hat{t}_r) = \sum_{h=1}^{H} N_h (\mu_{h,R=1} - \mu_h).$$

Writing $\mu_{h,R=0}$ as the mean of the survey variable in stratum h amongst those who do not respond and $\overline{R}_h$ as the rate of response in stratum h we may write

$$\mu_h = \overline{R}_h \mu_{h,R=1} + (1 - \overline{R}_h)\mu_{h,R=0}$$

and thus an alternative expression for the bias is

$$\text{bias}(\hat{t}_r) = \sum_{h=1}^{H} N_h (1 - \overline{R}_h)(\mu_{h,R=1} - \mu_{h,R=0}) \qquad (8.1)$$

Thus no bias arises if either there is no nonresponse ($\overline{R}_h = 1$) or if the respondents and nonrespondents share the same mean value of the survey variable within strata, which occurs when missingness is random within strata, that is when nonresponse is independent of the survey variable within strata. In general, however, this condition will not hold and nonresponse will lead to biased estimation of totals as well as of other population parameters.

8.3.3 Variance inflation

Within the setting again of Section 8.3.1, the variance of $\hat{t}_r$ will depend again on assumptions about the missing data mechanism. One simple assumption, which illustrates the variance impact of nonresponse, is that the respondents within stratum h form a simple random subsample of size m_h amongst the n_h units of the selected sample. In this case the variances before and after nonresponse respectively are

$$\begin{aligned}\operatorname{var}(\hat{t}) &= \sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}\\ \operatorname{var}(\hat{t}_r) &= \sum_{h=1}^{H} N_h^2\left(1-\frac{m_h}{N_h}\right)\frac{S_h^2}{m_h}\end{aligned} \tag{8.2}$$

where S_h^2 is the population variance in stratum h. Assuming an approximately uniform response rate across strata and ignoring the finite population corrections, the variance will be inflated by a factor roughly equal to the reciprocal of the response rate. Nonresponse in those strata with a high sampling fraction and especially in completely enumerated strata will tend to inflate the variance further.

8.3.4 Effects of confusing units outside the population with nonresponse

It will often be difficult to distinguish unit nonresponse from a unit which is outside the target population, for example because it has ceased to be active. If such a unit is treated as nonresponse then bias will usually arise. When estimating totals of variables such as production, a value of zero should be used whereas the treatment of nonresponse described in section 8.3.1 will effectively take the value as the stratum mean, biasing the estimate upwards. On the other hand, if a unit in the target population fails to respond and is wrongly treated as outside the target population then this will tend to lead to downward bias.

8.3.5 Effects of nonresponse on coherence

Many variables appearing in business surveys are subject to arithmetic constraints. For example, questions might be asked on capital expenditure under three headings as well as on total capital expenditure. There may be interest not only in the population totals A, B and C of the three specific types of capital expenditure but also in D = A+B+C, the total capital expenditure overall. However, item nonresponse may occur on different businesses, for different variables and so, if nonresponse is treated variable by variable as in 8.3.1, it is possible that the resulting estimates $\hat{\mathrm{A}}, \hat{\mathrm{B}}, \hat{\mathrm{C}}\,\text{and}\,\hat{\mathrm{D}}$ are not coherent, that is $\hat{\mathrm{D}} \neq \hat{\mathrm{A}}+\hat{\mathrm{B}}+\hat{\mathrm{C}}$. Many agencies may view such incoherence as undesirable, in particular because it may confuse users. Imputation provides one approach to dealing with this problem (see Section 8.6).

8.4 Quality measurement

8.4.1 Response rates

There are many response rates which may be calculated. Unit response rates may be calculated by size stratum and by industry stratum and may be weighted together across strata. Cumulative unit response rates may be calculated according to how many reminders have been issued. Unit nonresponse rates may be disaggregated by reason for nonresponse, noncontact, refusal etc. Item response rates may also be calculated for each survey variable.

The basic definition of a response rate is

$$\frac{\text{number of responding units}}{\text{number of eligible sample units}}$$

where an ‘eligible’ sample unit is one which is in the target population. The numerator is usually readily available. There may, however, be difficulties in determining the denominator because it may be difficult to decide whether sample units which do not respond are eligible. Some estimation of this number will generally be necessary, based for example on past estimates of ‘death rates’ of businesses.

Response rates have different uses, upon which the choice of rate will depend. One use is to monitor problems in data collection. For this purpose, it may be useful, for example, to record cumulative response rates over time following the initial issue of forms. Such evidence may be relevant, for example, to decisions about the timing and number of reminders.

The principal concern here is with the use of response rates for quality measurement. A basic problem is that the response rate is not directly related to the principal problem caused by nonresponse bias. It is, in principle, possible for nonresponse rates to be low and bias to be high and vice versa. Nevertheless, equation (8.1) does demonstrate an indirect relation between response rates and bias. If the response rates $\overline{R}_h$ within strata are high then the nonrespondents need to be much more different from the respondents to achieve the same level of bias as when the response rates $\overline{R}_h$ are much lower. High response rates might therefore be viewed as a form of **protection** against bias.

Comparing unit response rates between industry and size strata may be informative for quality control of data collection but these rates need summarising if an overall indicator of quality is to be determined. The way in which these rates should be summarised depends on the impact of nonresponse. A simple assumption is to suppose that the component $(\mu_{h,R=1} - \mu_{h,R=0})$ of bias in (8.1) is proportional to the mean $\overline{x}_h$ of a given auxiliary variable, such as employment, within stratum *h*. If it is also assumed for simplicity that the mean of the survey variable is proportional to $\overline{x}_h$ within strata we may approximate the relative bias by

$$\begin{aligned}\frac{\text{bias}(\hat{t}_r)}{\hat{t}_r} &\propto \frac{\sum_h N_h (1-\overline{R}_h)\overline{x}_h}{\sum_h N_h \overline{x}_h} \\ &= 1 - \frac{\sum_h t_{xh}\overline{R}_h}{\sum_h t_{xh}}\end{aligned}$$

where $t_{xh} = N_h \overline{x}_h$ is the stratum total of the auxiliary variable. Under these assumptions it seems appropriate to weight the stratum response rates $\overline{R}_h$ by the stratum totals t_{xh} if an overall measure of quality related to nonresponse bias is required. The weighted rate therefore takes the form

$$\text{weighted response rate} = \frac{\sum_{h=1}^{H} x_h \overline{R}_h}{\sum_{h=1}^{H} t_{xh}}, \tag{8.3}$$

where $\overline{R}_h$ is the response rate in stratum *h* and t_{xh} is the stratum total of an auxiliary variable judged to be proportional to the principal survey variables of interest. For example, if the auxiliary variable *x* is employment then this measure may be interpreted as the expected proportion of total employment in businesses which respond.

In order to reduce nonresponse bias it is common practice to devote greater resources to response chasing with the larger businesses. For example, in the Annual Business Inquiry, businesses of 200+ employment are targeted heavily. As a result the response rate $\overline{R}_h$ is higher in the larger size strata and the weighted response rate will be greater than an unweighted rate.

The sample version of formula (8.3) can be expressed as

$$\text{weighted response rate} = \frac{\sum_{k} w_k R_k}{\sum_{k} w_k} \tag{8.4}$$

where the sum is over sample units and the weight w_k for a sampled business in stratum *h* is $t_{xh}/n_h = w_h \overline{x}_h$, where n_h is the sample size in stratum *h* and $w_h = N_h/n_h$ is the expansion weight. Generalising the formula $w_h \overline{x}_h$, we may take

$$\begin{aligned} w_k &= \text{weight for sampled business } k \text{ in weighted response rate in (8.3)} \\ &= \text{estimation weight for business } k \times \text{size measure for business } k \end{aligned}$$

Such a weighted response rate reflects the relative importance of different sample units through both their weight in estimation and their size, assumed roughly proportional to the survey variable.

8.4.2 Measures based on follow-up data

Response rates are, however, unsatisfactory as measures of quality. Even if a lower response rate indicates the possibility of greater bias, response rates provide no information on how large that bias may be.

One approach to estimating nonresponse bias is to follow up nonrespondents (either unit or item nonrespondents) and collect the survey information from these businesses.

Two sources of bias can be addressed in this way. The most important source arises simply from the values missing due to nonresponse. These are collected in the follow-up survey. A second source arises because some assumed nonresponding units may in fact be ineligible and vice versa. Follow-up enables these two possibilities to be distinguished. Of course, complete response in the follow-up will rarely be achieved in practice and so the estimates of bias arising from follow-up data will themselves be subject to some error.

Most business surveys are subject to pressures for the early release of results. Sometimes this means that preliminary estimates are determined after an initial time period and final estimates are obtained after a longer period including perhaps further reminders. An estimate of the bias in the preliminary estimates is obtained simply from a difference between these estimates and the final estimates. This idea may be extended by collecting further data beyond the period upon which the final estimates are based. In this way the bias in the final estimates arising from nonresponse can be estimated.

In addition to extending the period available for data collection, other more intensive methods of follow-up can be used, in particular with different modes of data collection, such as the telephone and personal interview. Recognising the fact that fully successful follow-up is not only impractical but costly, selective follow-up strategies may be considered, focussed towards larger units which may be expected to make a greater contribution to the bias.

8.4.3 Comparison with external data sources and benchmarks

An alternative approach to estimating nonresponse bias is to make comparisons with external sources, such as other surveys, administrative sources or trade organisation data. National accounts sources may also provide benchmarks for comparison.

Two kinds of comparison are possible. First, comparison between overall estimates may be made. In this case differences between estimates may reflect not only nonresponse bias but also other sources of bias such as measurement error, and it may be difficult to disentangle these different sources. Moreover, differences between estimates may reflect bias in either the estimate of interest or in the comparative source and again it may sometimes be difficult to separate these effects. See the chapter on measurement errors (chapter 6) for an example of a comparison between a mail survey and an interviewer survey, where different rates of nonresponse arise.

A second kind of comparison may be undertaken when the survey respondents (and ideally nonrespondents) may be matched to records in the external source. The most obvious example is where the external source is the business register from which the sample was drawn. In this case comparisons may be made between respondents and other units in the external source with respect to variables available in that source. Another example is the comparison of survey responses with audited accounts, although these may only become available some time after the survey. Such comparisons may still be useful for assessing nonresponse bias even if the variables in the external source are subject to measurement error, so long as they are sufficiently correlated with the survey variables of interest.

8.4.4 Comparison of alternative adjusted point estimates

In sections 8.5 and 8.6 we consider weighting and imputation methods aimed at adjusting for nonresponse bias. These adjustment methods are based upon strong assumptions, in particular that nonresponse occurs at random given values of certain auxiliary variables (see section

8.2.2 for the definition of 'missing at random'). Departures from these assumptions may be expected to lead to biases in the adjusted estimators. Some assessment of bias may be made by comparing estimators based upon different assumptions, specifically using different choices of auxiliary variables.

In addition, the possibility of informative (non-ignorable) nonresponse (see section 8.2.2) may be considered. Alternative plausible models for informative nonresponse mechanisms might be specified and then the impact on estimation considered. Ways in which this might be done are discussed further in the chapter on model assumption errors (chapter 9). It may be possible to develop special estimation procedures under the specified informative nonresponse mechanisms as Copas & Li (1997) have done for certain modelling purposes. Alternatively, simulation-based procedures might be employed. Perhaps the most straightforward approach is to take a complete set of records from the sample data and treat this as if it is an 'artificial sample'. Next, missing values may be created in this artificial sample according to assumed nonresponse mechanisms (which may themselves have been arrived at by fitting models to the original data subject to nonresponse). Estimates may be computed from the new data according to the standard procedures employed in the survey and these estimates may be compared with estimates obtained from the full artificial sample. The process of creating missing values should preferably be repeated and the bias and variance of the estimator under the specified nonresponse mechanism estimated as in any simulation study.

8.5 Weighting adjustment

8.5.1 The basic method

The population total of a survey variable y_k is estimated by $\sum w_k y_k$, where the sum is across respondents. The basic idea is that each responding unit 'represents' w_k population units. The weight may be expressed as $w_k = w_{sk} w_{nrk}$, where w_{sk} is the sampling weight and w_{nrk} the nonresponse weight. Various methods may be used to construct the weights. In practice a single set of weights will usually be used for all survey variables. This is desirable not only for simplicity of computation but also to ensure that arithmetic relationships between variables (for example total capital expenditure is the sum of the components of capital expenditure) are preserved in the estimates. For this reason weighting, is the standard procedure used to adjust for unit nonresponse (which applies to all variables in a uniform way) but is usually unsuitable for item nonresponse, since different weights will be necessary for variables for which values are missing for different units.

8.5.2 Use of auxiliary information

In order to reduce nonresponse bias it is necessary to use auxiliary information about units which are not respondents. Two broad kinds of information may be used. First, certain information may be available on nonrespondents in the sample but not for other population units. One example arises in a monthly business survey when the sample consists of the same businesses each month. In this case information may be available on sample businesses in February, say, which may be used to weight for nonresponse in March. Such weighting is called *sample-based weighting*. Quantitative information on nonrespondents, such as reported values from the previous month in a monthly survey, is more likely to be used for imputation than for weighting. Categorical information, such as an industrial classification, might be used to define *response homogeneity groups* within which the nonresponse weights may be determined by the inverse response rates.

The second broad kind of information is that available on the whole population, most obviously information recorded on the business register. Weighting methods based on such information are called *population-based weighting*. The following two sections concern different methods of such population-based weighting.

8.5.3 Poststratification

This method is applicable when a classification of business is available which was not used for sampling. The classification partitions businesses into 'poststrata' *g*, where the number of businesses N_g within poststratum *g* is known. An example arises when the classification of businesses by industry or size is updated and considered to be more accurate than the original classification used for sampling (Hidiroglou *et al*., 1995). The poststratified estimator of a total takes the weighted form $\sum w_{sk} w_{nrk} y_k$ in section 8.5.1, where the nonresponse weight for all units in poststratum *g* is $w_{nrk} = N_g \big/ \hat{N}_g$, and $\hat{N}_g$ is obtained by summing the sample weights w_{sk} across responding units in poststratum *g*.

8.5.4 Regression estimation and calibration

Poststratification is a special case of regression estimation which itself is a special case of calibration estimation (Deville & Särndal, 1992; Lundström, 1997). Methods of ratio estimation used widely for business surveys are also special cases.

The simplest approach to handling unit nonresponse in these methods is to treat the respondents as the achieved sample with inclusion probabilities proportional to the sample inclusion probabilities. If the regression relationship between the survey variable and the auxiliary variables is the same for respondents and nonrespondents, the corresponding regression (or calibration) estimator will remove bias due to nonresponse (Hidiroglou *et al*., 1995, p.491). This is essentially the missing at random condition referred to earlier. Under departures from this assumption, regression estimation may still be useful for reducing nonresponse bias. A more complex approach involves first adjusting the sample inclusion probabilities by estimated nonresponse probabilities. Bethlehem (1988) argues that this adjustment may be expected to reduce bias.

8.5.5 Weighting and nonresponse errors

Weighting may be expected to affect both the bias and the variance arising from nonresponse. The aim is to remove nonresponse bias although, in practice, this is unlikely to be fully achieved. A comparison of alternative weighted estimators provides some idea of how bias may vary according to different assumptions. These assumptions will be of the form ‘missing at random given measured auxiliary variables’. These auxiliary variables might, for example, be those used to define response homogeneity groups in the sample, or to define poststrata for population weighting. A comparison of weighted estimators therefore represents a sensitivity analysis with respect to a limited set of assumptions.

Weighting will also generally affect the variance of the total survey errors in two ways. First, poststratification and more generally calibration weighting can act to reduce the variance if the auxiliary variables used help to predict the survey variables within strata. Second, variability in the weights can inflate the variance and this variance inflation tends to increase as the amount of auxiliary information increases (Nascimento Silva & Skinner, 1997).

8.5.6 Variance estimation

There exists a number of variance estimators in the presence of nonresponse. The simplest is to treat the nonresponse weights as fixed quantities for which variation between weights inflates the variance. This approach fails to allow for the reduction of variance achieved by population weighting. This variance reduction is allowed for by standard variance estimators for calibration estimation (for example Deville & Särndal, 1992). More complications arise if sample-based weighting is also involved. In this case, more complicated variance estimators are required, which include components both at the sample level and at the respondent level (Särndal & Swensson, 1987; Lundström, 1997). All of these estimators effectively make a missing at random assumption and thus do not allow for the possibility of informative nonresponse. See the chapter on model assumption errors (section 9.7) for further discussion of this case.

8.6 Imputation

8.6.1 Uses

Imputation is used generally for item nonresponse and, in particular, for allocating activity between the components, for example local units, of an enterprise when only aggregate values are reported. Imputation may also be used for unit nonresponse, especially for businesses in the completely enumerated stratum where previously reported values may be powerful predictors of missing values.

8.6.2 Deductive imputation and editing

The simplest form of imputation involves the use of logical relationships between variables and is usually performed as part of the editing process (Hidiroglou & Berthelot, 1986). For example if the total of non-negative variables is recorded as zero, then the values of these variables can be imputed as zero.

8.6.3 Last value imputation

For frequent (for example monthly) surveys a very simple imputation method is to use the most recently reported values.

8.6.4 Ratio and regression imputation

A simple modification of last value imputation is to scale this value with a ratio of estimates based on the current and previous values. Thus, if y_{tk} is the reported value of unit *k* at month *t* then y_{t+1k}, the value missing at month *t* + 1, might be imputed by

$$\hat{y}_{t+1k} = \frac{\bar{y}_{t+1r}}{\bar{y}_{tr}} y_{tk} \ .$$

Here the means $\bar{y}_{t+1r}$ and $\bar{y}_{tr}$ of reported values at months *t* + 1 and *t* respectively might be obtained from businesses in a similar industrial classification and size. Extreme values might be trimmed when calculating these means, to avoid outliers having excessive influence. This approach is particularly suited to variables which do not vary greatly over time.

More generally, a linear regression model $y = \sum x_p \beta_p$ might be fitted to the survey variable *y* with missing values, with the covariates x_p including previous values of the survey variable as well other variables, for example those on the business register. The imputed value may then be taken as the usual predictor $\sum x_p \hat{\beta}_p$, where $\hat{\beta}_p$ is the least-squares estimator of β_p. Business surveys tend to be well suited to such methods since strong correlations between variables are common.

8.6.5 Donor methods

Ratio and regression methods make efficient use of auxiliary information but are not suited to every application. They are difficult to apply to missing values in categorical variables and, since they are usually applied variable by variable, they may not preserve relationships between variables. In such circumstances, donor methods such as hot deck imputation may be useful. A donor unit is selected which is as similar as possible to the unit with missing values and the values from the donor are used to impute one or more missing values. Similarity may be measured for example according to size and industrial classification of the unit (Kovar & Whitridge, 1995).

8.6.6 Stochastic methods

A further problem with ratio and regression methods is that they tend to reduce the variation in the variables imputed. Often only national totals are of interest and this tendency will not be of concern. However, if distributional quantities are of interest, bias may arise. For example, if the proportion of businesses performing poorly according to some criterion is of interest, and the imputed values tend towards the centre of the performance distribution, this proportion may be underestimated. This problem may be addressed through the use of stochastic methods of imputation (Kovar & Whitridge, 1995). For example, the regression imputation $\sum x_p \hat{\beta}_p$ in section 8.6.4. might be replaced by the stochastic regression imputation $\sum x_p \hat{\beta}_p + e$, where e is a random residual, obtained by drawing a residual at random from those arising in the regression analysis used to obtain the $\hat{\beta}_p$.

8.6.7 Imputation and nonresponse errors

Like weighting, imputation may be expected to affect both the bias and variance arising from nonresponse.

Regarding bias, there are two broad considerations. The first one is the most obvious and applies equally to weighting. The success of imputation in removing bias for the estimation of characteristics of a given survey variable will depend on how well the imputation model captures the distribution of the missing value. Comparing the results of different imputation methods will provide some evidence on the size of such bias. A second, more subtle consideration is that imputation can introduce bias in estimates which depend on more than one variable if these variables are not fully controlled for in the imputation. Consider, for example, a variable which takes the following values for a business

	y	
December	1000	Reported
January	1050	Nonresponse (1000 imputed)
February	1100	Nonresponse (1000 imputed)
March	1150	Reported

Suppose that both the January and February values are missing and are each imputed by the last value 1000 (see section 8.6.3). Suppose that an estimate is required for the number of businesses which have changed y by over 100 from February to March. The above business will be erroneously classified in this category and imputation may lead to an upward bias in the estimation of this number. This could, in principle, have been avoided if the March figure had been used also to impute the February figure but, in practice, such ‘revisions’ are often viewed as undesirable.

Imputation may also be expected to have an impact on the variance of the estimator. In general, we may expect the variance to become $V_{\text{samp}} + V_{\text{imp}}$, where V_{samp} is the sampling variance which would have arisen in the absence of nonresponse and V_{imp} is the additional variance arising from imputation. The size of this term will depend on the form of imputation. The term V_{imp} will tend to be smaller for methods of ratio or regression imputation which are based on models with high predictive power. The term V_{imp} will tend to be larger for methods which have less predictive power, for example last value imputation, and for stochastic methods. Kovar & Whitridge (1995) suggest that imputation can inflate the variance by 2 to 10 percent in the case of a 5 percent nonresponse rate or by 10 to 50 percent in the case of 30 percent nonresponse.

8.6.8 Variance estimation

An important problem for quality measurement is that the variance impact of imputation is much harder to estimate than that of weighting. The simplest approach is to treat imputed values as real values and to use the usual estimators of sampling variance. Unfortunately, this will usually underestimate the variance because no account is taken of V_{imp}, the additional uncertainty arising from the fact that the imputed values will, in practice, not equal the true values. The degree of underestimation may be severe in business surveys, in particular because the usual estimators of sampling variance take no account of imputation error among large businesses in the completely enumerated strata.

Consider, for example, the use of a separate ratio estimator. The conventional variance estimator, treating the imputed values as real, takes the form

$$\sum_{h=1}^{H} N_h^2 \left(1 - m_h / N_h\right) s_h^2 / m_h \, , \tag{5}$$

by analogy with expression (2), where m_h is the number of units in stratum h for which data (including imputed values) are available, N_h is the corresponding population size and s_h^2 is the sample variance of the residuals (treating the imputed values as real). Assuming ratio imputation is employed using the same auxiliary variable as in the ratio estimator, the actual variance should be:

$$\sum_{h=1}^{H} N_h^2 \left(1 - m_h^* / N_h\right) S_h^2 / m_h^* \tag{6}$$

where m_h^* is the number of observations in stratum h *excluding* imputed values and S_h^2 is the variance of the residuals in the absence of item nonresponse. Assuming ratio imputation as above, each of the residuals in S_h^2 corresponding to an imputed value will be zero and $s_h^2 = \left(m_h^* - 1\right) S_h^2 / \left(m_h - 1\right)$. Thus the terms $\left(1 - m_h / N_h\right) s_h^2 / m_h$ in (5) tend to underestimate the corresponding terms $\left(1 - m_h^* / N_h\right) S_h^2 / m_h^*$ in (6) by a factor

$$\frac{\left(1 - m_h / N_h\right)\left(m_h^* - 1\right) m_h^*}{\left[\left(1 - m_h^* / N_h\right)\left(m_h - 1\right) m_h\right]} \approx \frac{\left(1 - m_h / N_h\right)}{\left(1 - m_h^* / N_h\right)} \left(\frac{m_h^*}{m_h}\right)^2$$

The amount of underestimation will tend to be large if either the sampling fraction m_h / N_h is large, especially for completely enumerated strata with $m_h / N_h = 1$, or if the fraction of imputed values $\left(1 - m_h^* / m_h\right)$ is large. A simple adjusted variance estimator takes the form

$$\sum_{h=1}^{H} \frac{N_h^2 \left(1 - m_h^* / N_h\right)\left(m_h - 1\right) s_h^2}{\left[m_h^* \left(m_h^* - 1\right)\right]}$$

and involves applying a correction to the standard variance estimator within each stratum. This estimator assumes that the same auxiliary variable is used for imputation as for estimation. This will often not be the case. An alternative approach to adjustment is to replace the imputed values by adjusted imputed values for the purpose of variance estimation. Suppose, for example, that imputed values are of the form $y_k^* = \hat{\beta} x_k$, where x_k is a previous value recorded for business *k* and $\hat{\beta}$ is a ratio. Then, for the purpose of variance estimation y_k^* might be replaced by $y_k^{**} = y_k^* + \varepsilon_k$, where ε_k is a randomly generated value from a normal distribution with mean zero and variance σ_k^2. The problem then is to choose the σ_k^2 in such a way that the standard variance estimator (treating the y_k^{**} as real values) is approximately unbiased for the total variance $V_{\text{samp}} + V_{\text{imp}}$. One approach is discussed by Rao (1996) in the context of jackknife variance estimation.

Särndal (1992) describes an approach which involves estimating the components V_{samp} and V_{imp} separately. A further approach is multiple imputation which involves creating multiple datasets with imputed values and comparing the estimates obtained from each (Rubin 1996, Fay 1996). None of these methods seem to have yet found their way into business survey practice in Europe, however, and the development and implementation of practical variance estimation methods remains an outstanding research problem.

9 Model Assumption Errors

David Draper & Russell Bowater[4], University of Bath

9.1 Introduction

The original goal of design-based analysis methods in survey sampling was "the development of a sampling theory that is model-free" (Cochran 1977). Even within classical design-based methods, however, the incorporation of auxiliary information through such techniques as ratio and regression estimation is essentially (if perhaps somewhat covertly) model-based. Today overtly model-based methods are commonly employed in business statistics, in the calculation of index formulae, in the use of benchmarking and seasonal adjustment (where model-based outlier detection and correction are crucial), and in estimation when no data for a sub-population are available (for example, enterprises that fall below a size threshold, as in cut-off sampling, or small-area estimation from aggregate data). Models are thus ubiquitous in the analysis of business survey data (see, for example, Särndal *et al.* 1992), and the assumptions they make must be critically reviewed with an eye to quantifying model assumption errors.

We have already encountered the use of models in several previous chapters; in particular, in section 2.3.2 we examined the idea of treating the population from which the sample at hand was drawn as itself a sample from a *superpopulation* specified by a model. An example of this idea that is relevant to model assumption errors came up in the discussion of quota sampling in section 4.3: if the population values y_j in the cells of the quota-sampling grid are assumed to be random variables with $\mathrm{E}_\xi\left(y_j\right) = \mu_h$, and $\mathrm{V}_\xi\left(y_j\right) = \sigma_h^2$, where h indexes the cell in the grid in which y_j is observed, then model-unbiased estimates both of the population total t ($\hat{t}$, say) and the variance of $\hat{t}$ are available and coincide with the usual design-unbiased estimates from stratified sampling. However, this is equivalent to the modelling assumption that the observed y_j values in the quota sample are stochastically indistinguishable from what one would obtain with simple random sampling (without replacement) from the cells in the grid, and there is no way to completely verify this assumption from the data. Errors in this model assumption could lead to a bias in the estimate of t whose magnitude and even direction are hard to quantify.

In the following sections we examine in turn the five leading areas in which model assumption errors appear crucial in business surveys: index formulae, benchmarking, seasonal adjustment, cut-off sampling, and coping with non-ignorable nonresponse. In the final section we offer some recommendations on best practice in the reporting of possible model assumption errors in business surveys.

[4] We are grateful to Ray Chambers (University of Southampton), Eva Elvers (Statistics Sweden) and Paul Smith (UK Office for National Statistics) for comments and references, and to Paul Smith for some suggested text fragments. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people responsible for any errors or omissions that may be present.

9.2 Index numbers

As noted by Jazairi (1982), an *index number* is a measure of the magnitude of a variable at one point relative to its value at another point. The variable in question is often either the price or the (sales) quantity (or volume) of a commodity. The “points” in question may be different times, or locations, or groups of households; we will focus here on time, measured in months. In the simplest form of this idea there are only two points in time being compared; one, say *t* (often the earlier time-point), is selected as the *reference* or *base month*, and the other, say t', is the *current month*.

Consider a set or *market basket, C,* of commodities $c_1,\ldots,c_m$ observed at *n* times, and let p_{it} and q_{it} be the price and volume, respectively, of commodity c_i at time *t*. The *money value* of c_i at time *t* is by definition simply the product $v_{it} \equiv p_{it}q_{it}$. The ratio $p_{it'}/p_{it}$ of the price of commodity c_i at time t' to its price at time *t* is the *price ratio*; the corresponding fraction $q_{it'}/q_{it}$ is the *volume ratio*. In attempting to measure how much the price of the market basket *C* has changed over time, an old (18th century) idea was simply to form the average $\frac{1}{m}\sum_{i=1}^{m}\frac{p_{it'}}{p_{it}}$ of the price ratios; in the 19th century the German economists Laspèyres and Paasche introduced a refinement of this idea which is still used today. The *Laspèyres price* and *volume indices*, respectively, are ratios of weighted sums of the form

$$LP_{tt'} = \frac{\sum_{i=1}^{m} p_{it'}q_{it}}{\sum_{i=1}^{m} p_{it}q_{it}}, \qquad LV_{tt'} = \frac{\sum_{i=1}^{m} q_{it'}p_{it}}{\sum_{i=1}^{m} q_{it}p_{it}}; \tag{9.1}$$

for example, the Laspèyres price index represents the ratio of the cost of the base month market basket at the current month prices to its cost at the prices of the base month. Similarly the *Paasche price* and *volume indices*, respectively, are

$$PP_{tt'} = \frac{\sum_{i=1}^{m} p_{it'}q_{it'}}{\sum_{i=1}^{m} p_{it}q_{it'}}, \qquad PV_{tt'} = \frac{\sum_{i=1}^{m} q_{it'}p_{it'}}{\sum_{i=1}^{m} q_{it}p_{it'}}; \tag{9.2}$$

thus the Paasche indices are similar to those of Laspèyres except that in Laspèyres’ weighted sums the weights are measured in the base month and Paasche's weights are those in the current month. With any given market basket, and base and current months, the Laspèyres and Paasche price indices will typically not agree (essentially for the same reason that the relative change of a quantity q_t from time *t* to t', $\left((q_{t'}-q_t)/q_t\right)$, does not coincide with the relative change from t' to *t*, $\left((q_{t'}-q_t)/q_{t'}\right)$); the *Fisher ideal index*, the geometric mean of the Laspèyres and Paasche formulae, is frequently used as a compromise. There are many variations on the idea illustrated here; Jazairi (1982) lists no less than 14 types of alternative index numbers.

A simple example of the role of model assumptions in the creation of index numbers arises from rewriting the Laspèyres price index as

$$LP_{tt'} = \frac{\sum_{i=1}^{m} p_{it'}q_{it}}{\sum_{i=1}^{m} p_{it}q_{it}} = \frac{\sum_{i=1}^{m}\left(\frac{p_{it'}}{p_{it}}\right)p_{it}q_{it}}{\sum_{i=1}^{m} v_{it}} = \frac{\sum_{i=1}^{m} v_{it}\left(\frac{p_{it'}}{p_{it}}\right)}{\sum_{i=1}^{m} v_{it}}, \tag{9.3}$$

thereby expressing this index as a weighted average of price ratios, using the values at time *t* as the weights. To produce $LP_{tt'}$ for time t', price ratios and values for time *t* are needed; in practice the values (often estimated from national accounts) might, for example, refer to the previous year and the price ratios might compare the current month with the previous December. At the time when the index is to be produced, reliable values for time *t* are often not yet available. It is then necessary to make an approximation, for example, to take values referring to an earlier year forward on the basis of some assumptions on growth rates. Any such assumptions will be model-based, either implicitly or explicitly, and the possibility of errors in the model assumptions ideally needs to be explored.

An example of an explicitly model-based approach to the construction of price and volume indices is given by the derivation of *best linear indices*. Theil (1960), the originator of this method, assumes that the prices of the *m* commodities move proportionately, apart from random fluctuations. As noted by Fisk (1977), one way to express this assumption is through the model

$$\sum_{i=1}^{m} w_{it'}\frac{p_{it}}{p_{it'}} = \sum_{i=1}^{m} v_{it'}p_{it} + e_{tt'}, \tag{9.4}$$

in which typically "$w_{it'} = p_{it'}v_{it'}$ is the average money value recorded as spent by a sample group of households on commodity *i* in time period t', and $p_{it'}/p_{it}$ is the price ratio for commodity *i* obtained from an independent source, usually a survey of prices in retail outlets." Here $e_{tt'}$ is treated as a stochastic error term assumed to have mean zero, although Fisk notes that "in practice non-sampling errors may prove more important than sampling errors and $e_{tt'}$ may contain a bias component which is not necessarily constant for all pairs (t,t')." To construct the price and volume indices for *m* commodities over *n* time periods one may form the $n \times m$ price and quantity matrices *P* and *Q*, define the money value matrix $M = PQ^{\mathrm{T}}$, and obtain the best linear price and volume indices *p* and *q* by unweighted least squares, as the vectors that minimise the sum of squares of the elements of the residual matrix $R = M - pq^{\mathrm{T}}$. In Section 9.8 we discuss how to assess the effects of errors in the assumptions underlying models such as (9.4).

9.3 Benchmarking

A good definition of this topic is given by Cholette & Dagum (1994):

> "Benchmarking situations arise whenever two (or more) sources of data are available for the same *target variable* with different frequencies, for example, monthly versus annually, or monthly versus quarterly. Generally, the two sources of data do not agree; for example, the annual sums of monthly measurements of a variable are not equal to the corresponding annual measurements. Furthermore, one source of data, typically the less frequent, is more reliable than the other, because it originates from a census, exhaustive administration records, or a larger sample. The more reliable measurements are considered as *benchmarks*. Traditionally, benchmarking has consisted of adjusting the less reliable series to make it consistent with the benchmarks. Benchmarking, however, can be defined more broadly as the process of optimally combining two sources of measurements, in order to achieve improved estimates of the series under investigation. Under such a definition, bench-marks are treated as auxiliary observations.
>
> A typical example of benchmarking is the following. In Statistics Canada, the monthly estimates of wages and salaries originate from the Survey of Employment, Payrolls, and Hours, whereas the annual benchmark measurements of the same variables originate from exhaustive administrative records, namely the income tax forms filed by Canadians and compiled by Revenue Canada. Benchmarking adjusts the monthly data so that they conform to the benchmarks and preserve the original month-to-month movement as much as possible."

Continuing the context of the last paragraph in this quote, in this section we take the less-frequent series – the benchmarks – to be annual and the more frequent series to be monthly, and we use wages and salaries as the outcome variable of interest.

For most of the past 25 years, most statistical agencies worldwide have performed benchmarking using one variation or another of a method proposed by Denton (1971). In this method, which was not originally based on a statistical model for the two time series, the monthly series is required to exactly match the benchmarks, which are regarded as *binding*, but as much as possible of the month-to-month movement of the original less-reliable series is preserved. More recently, explicitly model-based methods have emerged – for example, those of Cholette & Dagum (1994, hereafter CD) and Durbin & Quenneville (1997), following on from work of Hillmer & Trabelsi (1987) – which attempt to generalize the Denton approach to increase the realism of the benchmarking.

CD observe that survey errors (of the type likely to affect the monthly data) are often heteroskedastic and autocorrelated, and the survey may be biased due to factors such as non-ignorable nonresponse (see section 9.7) and frame deterioration over time. They propose an improvement to the Denton method based on a regression model that takes account of these factors. Their model is

$$\begin{aligned} \hat{t}_t &= a + \theta_t + e_t, \quad t = 1,\ldots,T; \\ z_m &= \sum_{t\in m} \theta_t + w_m, \quad m = 1,\ldots,M. \end{aligned} \tag{9.5}$$

Here $\{\hat{t}_t, 1 \le t \le T\}$ is the series of monthly measurements, decomposed into the sum of (i) a bias term *a*; (ii) the underlying "true" (unobserved) values of the wages and salaries series θ_t; and (iii) survey errors e_t, assumed to satisfy $\mathrm{E}(e_t) = \mathrm{E}(e_t e_{t-k}) = 0$ for all *t* and *k*. $\{z_m, 1 \le m \le M\}$ is the series of annual benchmarks, potentially subject to the errors w_m which satisfy $\mathrm{E}(w_m) = \mathrm{E}(w_m w_{m-k}) = 0$ for all *m* and *k* (the e_t and w_m are taken to be mutually independent). If the benchmarks are thought not to be subject to error then the w_m may all be taken to be zero; in this case the z_m series is binding.

Model (9.5) may be written in the familiar regression form

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathrm{E}(\mathbf{u}) = \mathbf{0}, \quad \mathrm{E}(\mathbf{u}\mathbf{u}^{\mathrm{T}}) = \mathbf{V}, \tag{9.6}$$

where the $\boldsymbol{\beta}$ vector includes both *a* and the vector $\boldsymbol{\theta}$ of true values. Autocorrelated errors e_t in the monthly series can be accommodated in this method by assuming that the e_t follow a stationary *ARMA*(*p*, *q*) model and computing the (estimated) covariance matrix **V** in (9.6) in terms of the estimated parameters of the *ARMA* model. Weighted least squares, taking the resulting matrix $\hat{\mathbf{V}}$ as known, then produces the usual estimate $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{z}$, from which complicated matrix expressions for $\hat{a}$ and $\hat{\boldsymbol{\theta}}$ (which we omit) may be deduced. Heteroscedasticity in the e_t may also be handled by writing $e_t = k_t e_t^*$, where the k_t, the standard deviations of the monthly series errors, are allowed to vary over time, and letting the e_t^* (not the e_t) follow an *ARMA* model. CD show that Denton-type methods for benchmarking are a special case of this regression framework, and they also demonstrate that their approach is more efficient than Denton adjustment under a variety of time series models for the e_t.

Durbin & Quenneville (1997, hereafter DQ) take a different approach to the construction of optimal benchmarking estimates, based on state-space structural time series models. Their approach assumes an additive error structure for the annual series, but can handle either additive or multiplicative behaviour of the monthly series. In the case of additive monthly errors, for example, DQ assume that the monthly series $\hat{t}_t$ follows the model

$$\hat{t}_t = \eta_t + k_t u_t, \qquad t = 1, \ldots, T, \tag{9.7}$$

where the η_t are underlying true values, the k_t are standard deviations of the survey errors, and the u_t are taken to be realisations of a unit-variance stationary *ARMA*(*p*,*q*) model. *p*, *q*, and the k_t are assumed known from substantive knowledge of the survey. They further assume that the annual benchmarks $\mathbf{z} = (z_1, \ldots, z_M)^{\mathrm{T}}$ are related to $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)^{\mathrm{T}}$ through the regression model

$$\mathbf{z} = \mathbf{L}\boldsymbol{\eta} + \mathbf{e}, \qquad \mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_e), \tag{9.8}$$

where the matrices $\mathbf{L}$ and $\mathbf{\Sigma}_e$ are again assumed known. As with the approach of Cholette & Dagum (1994), when the error vector $\mathbf{e}$ is assumed to be zero the benchmarks are binding. The state-space character of DQ's model enters through the assumption that

$$\eta_t = \mu_t + \gamma_t + \sum_{j=1}^{k} \delta_{jt} w_{jt} + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right), \tag{9.9}$$

where μ_t accounts for any trend that may be present, γ_t models the seasonal component (if any), and $\sum_{j=1}^{k} \delta_{jt} w_{jt}$ is the *trading-day* adjustment. Many models are available for the trend and seasonal components (for example, Harvey 1989); DQ use

$$\begin{aligned} \mu_t &= 2\mu_{t-1} - \mu_{t-2} + \zeta_t, \quad & \zeta_t \sim N\left(0, \sigma_\zeta^2\right) \\ \gamma_t &= -\sum_{j=1}^{11} \gamma_{t-j} + \omega_t, \quad & \omega_t \sim N\left(0, \sigma_\omega^2\right) \end{aligned} \tag{9.10}$$

The first of these equations yields a constant linear trend if $\sigma_\zeta^2 = 0$ but otherwise adapts to a time-varying slope; the second forces a constant seasonal pattern if $\sigma_\omega^2 = 0$ but permits this pattern to vary over time otherwise. DQ's model is completed with the assumption that the coefficients in the trading-day adjustment follow the relation

$$\delta_{jt} = \delta_{j,t-1} + \varsigma_{jt}, \qquad \varsigma_{jt} \sim N\left(0, \sigma_\varsigma^2\right); \tag{9.11}$$

here once again, constant coefficients are obtained by setting $\sigma_\varsigma^2 = 0$, but time-varying coefficients may be accommodated otherwise. All of the error series – ε_t, ζ_t, ω_t and ς_t – are assumed to be jointly independent of each other and of u_t. DQ use standard *Kalman filtering and smoothing* (KFS) methods (see, for example, Harvey 1989) to fit this model.

It is clear from equations (9.5-9.11) that benchmarking methods in current use or recently proposed are based on models with strong structural and distributional assumptions. Effects of errors in modelling assumptions like those in benchmarking are discussed in section 9.8.

9.4 Seasonal adjustment

Many business time series exhibit seasonal variation, typically annual in pattern when the series is observed monthly. Harvey (1989) defines seasonal trend as “that part of the series which, when extrapolated, repeats itself over any one-year time period and averages out to zero over such a time period.” Since such trend “contains no information on the general direction of the series, either in the long run or the short run,” seasonality is usually dealt with by estimating it, subtracting out the estimate, and studying the properties of the resulting seasonally-adjusted series. Simple *ad hoc* estimates can readily be conceived – for example, as Chatfield (1996) notes, “For series showing little trend, it is usually adequate to estimate the seasonal effect for a particular period (for example, January) by finding the average of each January observation [in the observed time series] minus the corresponding yearly average” when the seasonal component is thought to be at least roughly additive in character – but in practice more complicated model-based methods are typically employed.

For example, the UK Office for National Statistics (ONS) uses the computer program *X11-ARIMA* for almost all of its seasonal adjustment. *X11-ARIMA* involves the choice of an appropriate *autoregressive integrated moving average* (ARIMA) model (for example, Box, Jenkins & Reinsel 1994) for forecasting observations at both ends of a finite series, and this augmented series is then passed through a series of *Henderson filters* (for example, Kenny & Durbin 1982) involved in (a) outlier detection and removal/down-weighting, (b) choice of an appropriate filter for seasonal adjustment (generally based on the irregular-to-cyclic (IC) ratio) and (c) the adjustment itself. (Henderson filters are smoothing techniques based on moving averages which “aim to follow a cubic polynomial trend without distortion” (Chatfield 1996)).

Researchers at the US Census Bureau (Findley *et al.* 1998) have recently released *X12-ARIMA*, a superset of *X11-ARIMA* based on *regARIMA* modeling and intended to improve on the older software in a number of ways. As noted by these authors,

> “The basic seasonal-adjustment procedure of *X11-ARIMA* and [its predecessor] *X-11* decomposes a monthly or quarterly time series into a product of (estimates of) three components: a *trend* component, a *seasonal* component, and a residual component called the *irregular* component. Such a *multiplicative decomposition* is usually appropriate for series of positive values (sales, shipments, exports, etc) in which the size of the seasonal component increases with the level of the series, a characteristic of most seasonal macroeconomic time series. Under the multiplicative decomposition, the *seasonally adjusted series* is obtained by dividing the original series by the estimated seasonal component. ...
>
> Given a time series Y_t to be modeled, it is often necessary to take a nonlinear transformation of the series, $y_t = f_t(Y_t)$, to obtain a series that can be adequately fit by a regARIMA model. For example, if Y_t is a positive-valued series with seasonal movements proportional to the level of the series, one would usually take logarithms or, more generally, [work with]
>
> $$y_t = \log\left(\frac{Y_t}{d_t}\right) = \log Y_t - \log d_t, \qquad (9.12)$$
>
> where d_t is some appropriate sequence of divisors. ...
>
> Let *B* denote the backshift operator, $By_t = y_{t-1}$. *X12-ARIMA* can estimate regARIMA models of order $(p,d,q)(P,D,Q)_s$ for y_t. These are models of the form
>
> $$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D\left(y_t - \sum_{i=1}^{r}\beta_i x_{it}\right) = \theta_q(B)\Theta_Q(B^s)a_t, \qquad (9.13)$$
>
> where *s* is the length of the seasonal period [(typically *s* = 12)].”

Here a_t is a white-noise IID series with mean 0 and standard deviation σ_a, the x_{it} are *r* time series thought to be predictive of y_t, and $\phi_p(z)$, $\Phi_P(z)$, $\theta_q(z)$ and $\Theta_Q(z)$ are polynomials of degree *p*, *P*, *q*, and *Q*, respectively. In the usual way with ARIMA models, *p* and *P* are the orders of the autoregressive parts of the non-seasonal and seasonal models, *q* and *Q* are the orders of the moving average parts, and *d* and *D* are the numbers of times the non-seasonal and seasonal parts of the series need to be differenced to achieve stationarity. The same definitions apply to *X11-ARIMA*.

The default ARIMA model used by the ONS is $(0,1,1)(0,1,1)_{12}$ (the first model tested by X11-ARIMA, and accepted in the majority of cases, although other models are used too). A different default model, $(0,2,1)(0,1,1)_{12}$ is used in trend estimation. The selection of a Henderson filter for the main seasonal adjustment part is automatic based on the IC ratio, with choice between a 13-term and 23-term moving average for monthly series. There are several levels of adjustment for more or less severe outlier removal, in each case with the most atypical observations completely replaced by an estimate consistent with the model, and with the weight of other outliers reduced in the seasonal adjustment. The analyst can choose certain data points to be marked manually as outliers, but this is more often done through prior adjustments in which the reason for an unusual observation is noted (for example, a strike action). These prior adjustments can be temporary (unusual residuals that feed through to seasonally adjusted series) or permanent (adjusted data also used in outputs).

The principal assumptions which affect the ONS seasonal adjustment method and hence the final data can thus be summarised as follows:

(i) use of X11-ARIMA over any other seasonal adjustment software, with implicit reliance on Henderson filters in all cases (rather than, say, *Kalman filters* (for example, Abraham & Ledolter 1983; see below) or straightforward Box-Jenkins-style ARIMA modelling);

(ii) choice of level of outlier detection/treatment;

(iii) use and extent of permanent/temporary prior adjustments; and

(iv) the details of the ARIMA model used for forecasting beyond the ends of the finite input series.

The effects of errors in modelling assumptions such as these will be considered in section 9.8. The problem is particularly difficult because seasonal adjustment is an attempt to estimate a *counter-factual* outcome – namely, what values the time series undergoing seasonal adjustment *would have* exhibited had there been no seasonal effect – so that no “gold standard” (true) values are available for comparison.

A leading alternative to the *X11(X12)-ARIMA* approach to seasonal adjustment is found in the programs *TRAMO* and *SEATS* developed by Gomez & Maravall (1994a, b) at the Bank of Spain and now in widespread use throughout Europe. *TRAMO* (Time series Regression with ARIMA noise, Missing Observations, and Outliers) is a regARIMA model-based method which estimates missing data, identifies and downweights four kinds of outliers, and copes with special circumstances such as holiday and calendar effects. *TRAMO* can be used as a pre-processor to *SEATS* (Signal Extraction in ARIMA Time Series), which uses minimum mean-squared error methods to decompose the series into trend, seasonality, cycle, and irregular components. Findley *et al.* (1998) observe that the *TRAMO-SEATS* procedure "is equivalent to the modified Kalman-filter of Kohn & Ansley (1986), which extends the approach proposed by Jones (1980) to the case of models with differencing and missing data in the first $d+sD$ time points." These authors found in a comparison of *X12-ARIMA* and *TRAMO-SEATS* on data in which observations had been deliberately set aside and marked missing that "the estimates of the missing values from both procedures were always close to each other, and were also usually quite close to the value of the deleted datum (< 2% error)." Maravall (1998), in his discussion of the Findley *et al.* paper, asserts that *TRAMO-SEATS* is superior to *X12-ARIMA* in some respects, but Findley *et al.* demonstrate in their rejoinder that the two approaches often produce similar results (see Eurostat 1998b for additional comparisons).

9.5 Cut-off sampling

Continuing the discussion of cut-off sampling in chapter 4, consider a population of *N* companies and let x_j be register employment at some fixed time point of interest, sorted from largest to smallest, with y_j the corresponding turnover values. The total turnover $t_y = \sum_{j=1}^{N} y_j$ is to be estimated. Let $t_x = \sum_{j=1}^{N} x_j$.

In cut-off sampling t_x will be known, but only (at most) the first *k* of the y_j will be observed, where (in one leading application of the method) *k* is the smallest integer such that

$$t_{xk} = \sum_{j=1}^{k} x_j \geq (1-\varepsilon)t_x \qquad (9.14)$$

for some $0 < \varepsilon < 1$ (typically on the order of 0.05-0.2). With this approach complete enumeration of all of the $\{y_j, j \leq k\}$ may be undertaken, or a sample may be chosen; we focus here on the former case.

Having identified *k*, it is useful to define $t_{yk} = \sum_{j=1}^{k} y_j$ and to decompose t_y into the sum $t_{yk} + t_{yk}^*$, where $t_{yk}^* = \sum_{j=k+1}^{N} y_j$. In section 4.5.1 we examined the approach to estimating t_y based on ignoring t_{yk}^* (in effect, estimating it by 0); here we consider the effects of model assumption errors on attempts to estimate t_{yk}^* .

Perhaps the simplest estimate is obtained by defining $t_{xk}^* = \sum_{j=k+1}^{N} x_j$ and making the (unverifiable) assumption that

$$\frac{t_{yk}^*}{t_{xk}^*}=\frac{1\big/(N-k)\sum_{j=k+1}^{N} y_j}{1\big/(N-k)\sum_{j=k+1}^{N} x_j}\equiv\frac{1\big/k\sum_{j=1}^{k} y_j}{1\big/k\sum_{j=1}^{k} x_j}=\frac{t_{yk}}{t_{xk}}. \tag{9.15}$$

If (9.15) were true then t_y could be estimated by

$$\left(\hat{t}_y\right)_{ratio}=t_{yk}+\hat{t}_{yk}^*=t_{yk}+\frac{t_{yk}t_{xk}^*}{t_{xk}}=t_{yk}\frac{t_{xk}+t_{xk}^*}{t_{xk}}=\frac{t_{yk}}{t_{xk}}t_x, \tag{9.16}$$

which is recognisable as a simple ratio estimator.

For example, in one of the simulated populations based on the 1996 *ABI* (Annual Business Inquiry) data described in Section 4.5, there were *N* = 2,453 companies in the population, with total register employment across all *N* companies of $t_x = 169{,}013$ and true total turnover $t_y = 21{,}739{,}196$. Using an ε value of 0.2 (cutting off 20% of the employees, so to speak) yields $k = 699$ companies in the sample; for these companies $t_{xk}=\sum_{j=1}^{k} x_j = 135{,}241$ and $t_{yk}=\sum_{j=1}^{k} y_j = 18{,}884{,}196$. In this case, ignoring t_{yk}^* altogether would yield an estimate of $\hat{t}_y = t_{yk}$, which is biased low by $(21{,}739{,}196-18{,}884{,}196)/21{,}739{,}196 = 13.1\%$. If instead assumption (9.15) were made, the resulting ratio estimate would be $\left(\hat{t}_y\right)_{ratio}=\frac{18{,}884{,}196}{135{,}241}169{,}013 = 23{,}599{,}904$, which is biased high by 8.6%.

In this example the ratio estimator achieved a bias reduction of $(13.1-8.6)/13.1 = 35\%$, but larger improvements are possible. To see why requires motivating ratio estimation from a model-based perspective and looking for model assumption errors. It can be shown (see Cochran 1977 or Särndal *et al.* 1992) that if the *N* population values $\left(x_j, y_j\right)$ are themselves assumed to be a random sample from a superpopulation in which

$$y_j = \beta x_j + e_j, \tag{9.17}$$

where the e_j are independent of the x_j and satisfy $\mathrm{E}\left(e_j\right)=0,\quad \mathrm{V}\left(e_j\right)=\sigma^2 x_j$, then $\left(\hat{t}_y\right)_{ratio}$ is best linear unbiased for t_y with any sample, random or not, selected solely according to the values of the x_j. Thus, in this particular sense, the "model underlying" $\left(\hat{t}_y\right)_{ratio}$ (or, at least, a leading situation in which $\left(\hat{t}_y\right)_{ratio}$ would be expected to perform well) is a linear regression through the origin of the y_j on the x_j, in which the variance of y_j is proportional to x_j.

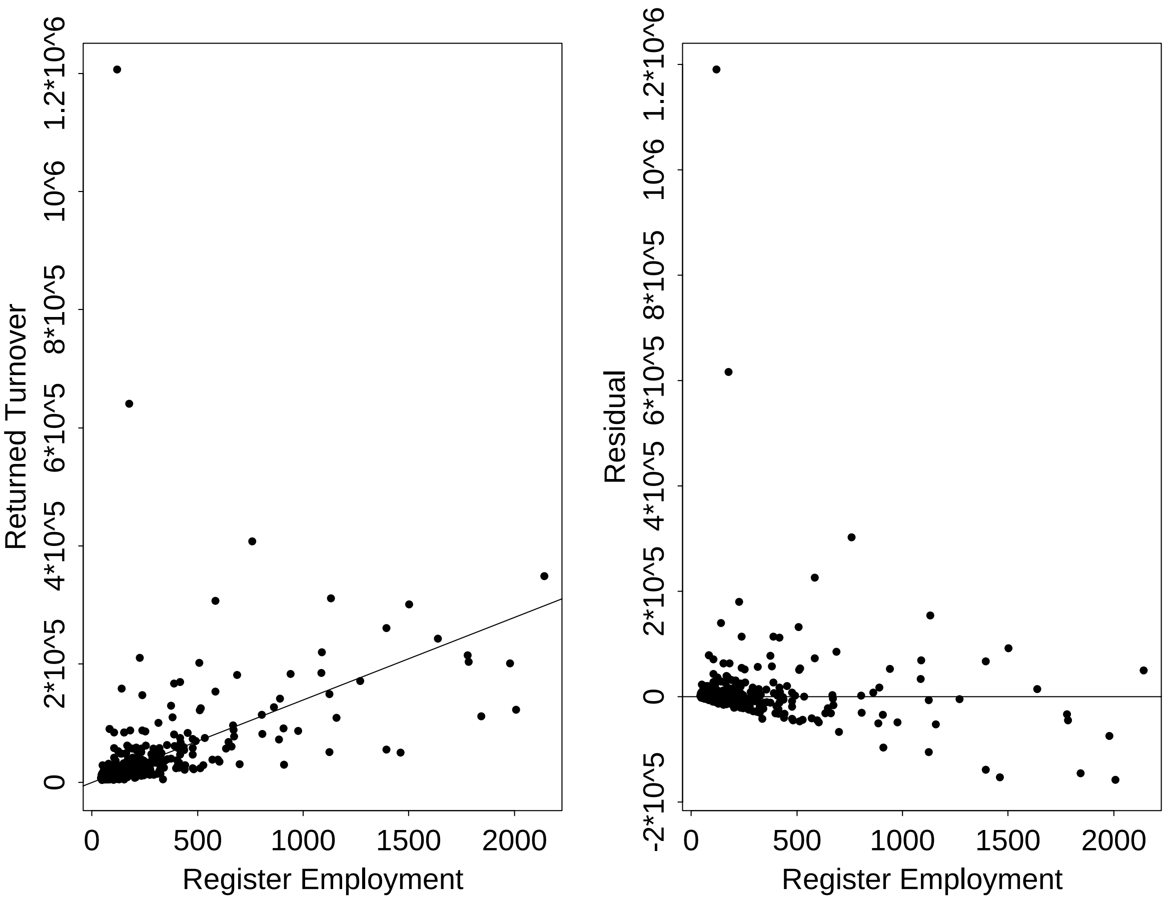


Figure 9.1 Scatterplot (left panel) and residual plot (right panel) from fitting model (9.17) to the ABI simulated population.

Standard statistical/econometric model-checking methods such as scatterplots and residual plots are helpful in evaluating the fit of model (9.17). The left panel of Figure 9.1 is a scatterplot of returned turnover against register employment for the 699 sampled companies in the ABI example above, with the fitted line $\hat{y}_j = \frac{t_{yk}}{t_{xk}} x_j$ from the ratio estimation model superimposed. It is evident from the sharply non-elliptical shape of these plots that least squares − even weighted least squares − is not making the best use of the bivariate data (x_j, y_j), and it is also clear that the estimated slope is quite possibly being driven by a small number of points with large register employment values. A standard remedy for this is to trim a small fraction, say 100γ%, of points with the largest x_j before estimating the slope, where $\gamma$is perhaps in the range 0.01–0.10. Denote the resulting population total estimate by $(\hat{t}_y)_{ratio}^{trim}$.

Another standard approach to estimating t_{yk}^* arises from relaxing the assumption of a zero intercept in fitting a linear model to the (x_j, y_j) pairs. Figure 9.1 does appear to indicate some sort of heteroskedasticity (that is, $\mathrm{V}(y_j)$ is not constant with varying x_j), but the strong clustering of the points near the origin makes it difficult to see what form the variance function should take. Assuming constant variance as a starting point amounts to fitting the model

$$y_j = \beta_0 + \beta_1 x_j + e_j \qquad \mathrm{E}(e_j) = 0, \quad \mathrm{V}(e_j) = \sigma^2 \tag{9.18}$$

by ordinary (unweighted) least squares (OLS), leading to the following estimate of total turnover:

$$\left(\hat{t}_y\right)_{reg} = t_{yk} + \sum_{j=k+1}^{N} \left(\hat{\beta}_0 + \hat{\beta}_1 x_j\right). \tag{9.19}$$

As with ratio estimation, it is sensible to trim the $100\gamma\%$ of points with the largest x_j before estimating the slope, yielding the estimator $\left(\hat{t}_y\right)_{reg}^{trim}$. With or without trimming, regression (rather than ratio) estimation may be a poor choice in cut-off sampling, leading to even more biased estimates than those produced by the untrimmed ratio method: with the example data given above, for instance, $\left(\hat{\beta}_0, \hat{\beta}_1\right) = (2805.7,\ 125.1)$, leading to $\left(\hat{t}_y\right)_{\mathrm{reg}} = 28{,}031{,}374$, which is biased high by 28.9%. What is worse, this method does not even guarantee positive predicted turnover values (attempting to respond to any heteroscedasticity that may be present, by using weighted least squares with a free intercept parameter and with y_j on the raw scale, may also fall victim to negative total turnover estimates).

The natural data-analytic solution to these problems is to find a scale for the (x_j, y_j) values on which OLS performs well (and on which the estimated total turnover cannot be negative). The vigorous bunching up of the points in the lower left corner of the scatterplot suggests a logarithmic transformation for both variables. So let $y'_j = \log(y_j)$ and $x'_j = \log(x_j)$, and regress y'_j on x'_j using OLS, obtaining intercept and slope values (on the log scale) $\hat{\beta}'_0$ and $\hat{\beta}'_1$, respectively. Then solving the equation

$$\log\left(\hat{y}_j\right) \cong \left(\widehat{\log y_j}\right) = \hat{\beta}'_0 + \hat{\beta}'_1 \log\left(x_j\right) \tag{9.20}$$

for $\hat{y}_j$ yields a log-log regression estimate of total turnover:

$$\left(\hat{t}_y\right)_{reg(\log)} = t_{yk} + \sum_{j=k+1}^{N} e^{\hat{\beta}'_0} x_j^{\hat{\beta}'_1}. \tag{9.21}$$

(This estimate is biased due to the nonlinear nature of the log transformation and could potentially benefit from bias adjustment, but the bias is small in this example, as Table 9.1 will demonstrate.)

Figure 9.2 parallels Figure 9.1, this time on the log-log scale. With this transformation the point-cloud is nicely elliptical (except for the left-truncation caused by cutting off the smallest companies), and OLS should perform efficiently. With the example data given here,

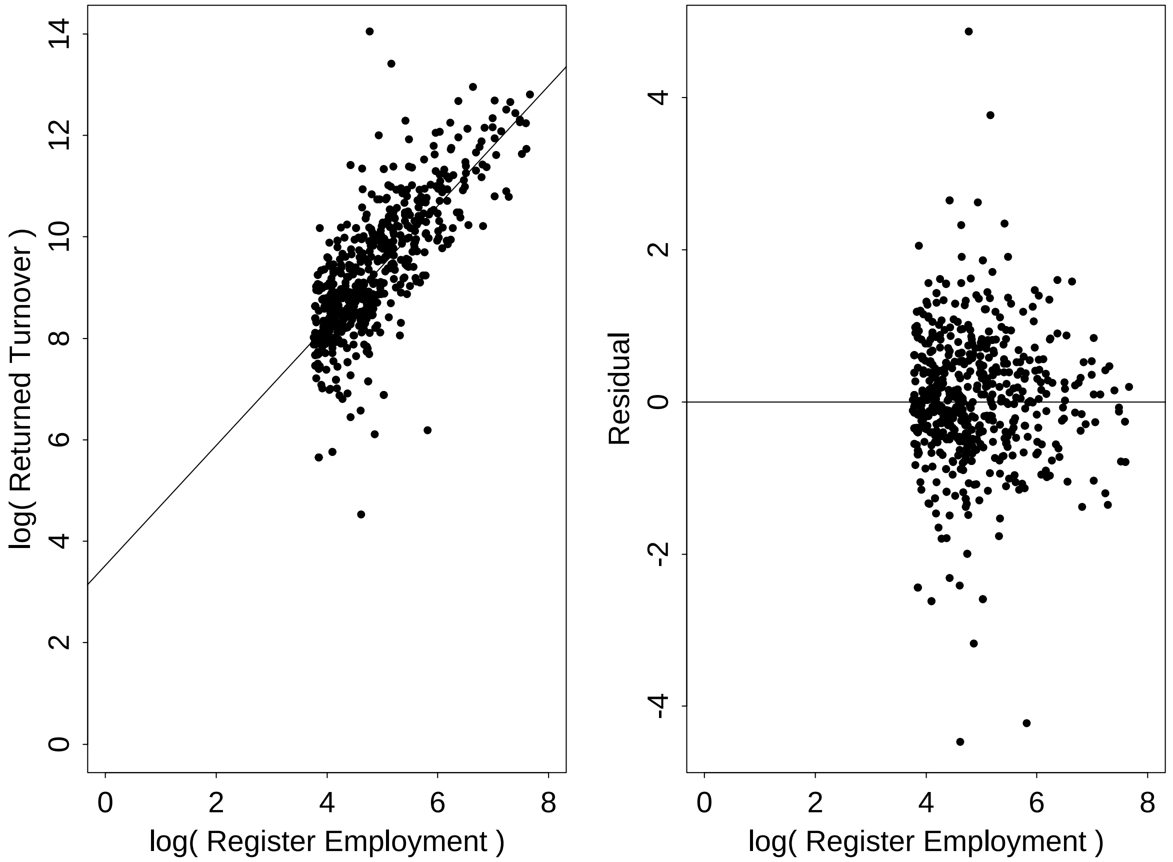


Figure 9.2 Scatterplot (left panel) and residual plot (right panel) from fitting a linear model on the log-log scale to the ABI simulated population.

the results are $\left(\hat{\beta}_0', \hat{\beta}_1'\right) = (3.53, 1.18)$ and $\left(\hat{t}_y\right)_{reg(log)} = 20{,}885{,}766$, which differs from the true population value by only 3.9% on the low side. On the log-log scale register employment and returned turnover have a sample correlation of +0.84 (the corresponding figure on the raw-raw scale is +0.56), and regression estimation on this scale can make effective use of this relationship. There is no need to trim any points with this approach, because the log transformation has removed the high-leverage nature of the companies with large register employment.

Table 9.1 presents the results of a simulation comparing the three total turnover estimators $\left(\hat{t}_y\right)_{ratio}$, $\left(\hat{t}_y\right)_{reg(raw)}$ and $\left(\hat{t}_y\right)_{reg(log)}$. As in Section 4.5 we repeatedly (100 times) (a) drew a sample of size 2,453 (the ABI extract sample size in 1995) with replacement from the ABI data but with unequal selection probabilities determined by the sampling weights, to create a pseudo-population reflecting the actual distribution of UK companies, and (b) used the register employment variable in this population to cut off the lower 100ε% of the companies (by cumulative employee numbers); but this time we (c) estimated the total returned turnover with each of the three estimators studied here, varying the trimming fraction γ in the case of the first two estimators from 0.01 to 0.10.

ε	Mean relative bias (%)					Optimal trim fraction	
	Ratio	Ratio trimmed	Regression raw	Regression raw trimmed	Regression log	Ratio	Regression raw
0.20	+9.1	+3.7	+32.7	-5.0	-2.8	0.10	0.10
0.15	+6.8	+0.2	+17.6	-7.2	-2.3	0.07	0.08
0.10	+4.4	+1.3	+8.0	-5.8	-1.6	0.03	0.06
0.05	+2.1	-1.4	+2.6	-3.0	-0.9	0.02	0.04

Table 9.1 Simulation results with the 1996 ABI data. The mean value of the true population total turnover across the 100 simulated replications was 26,650,310.

From the table it is evident that

- with this type of population the untrimmed ratio estimator is biased high by an unacceptably large margin for all but the smallest values of ε – in fact, comparing these results with those in Table 4.9, ratio estimation without trimming is almost as bad as ignoring the cut-off units altogether. However, after trimming, the ratio estimation approach performs very well, with relative errors of less than 0-4%;
- untrimmed regression estimation on the raw scale is even worse than untrimmed ratio estimation, overstating the true population total by up to 33% as a function of cut-off fraction. Trimming helps, but not enough to make the method viable with ABI-type data; and
- regression estimation on the log-log scale (without any need to search for an optimal trimming fraction) performs very well, yielding discrepancies between estimated and true totals on the order of only 1-3% of the truth.

This example has illustrated the value of both (a) standard statistical model-checking methods such as the examination of residuals and (b) sensitivity analysis, exploring a variety of models (in this case, (9.17), (9.18), and the log-log model leading to (9.21)) to observe the effects of model assumptions on the quantities of direct interest.

9.6 Small domains of estimation

Most of the discussion in this report has so far focused on the estimation of a total or mean for the entire population of interest (an exception is section 3.1). In many business surveys, however, there is also interest in estimating the total or mean in subsets, or *domains*, of the population. Sometimes these domains are defined by variables along which stratification has taken place in the survey design. In such cases it is often possible to over-sample rare subgroups (or *small domains*) to obtain accurate domain-specific estimates, without sacrificing much accuracy in the overall population estimates (see, for example, Cochran 1977). In other cases, however, the domains are too numerous for this strategy to work effectively. A classic example occurs when a survey is carried out fairly sparsely over a wide geographic region, but it is still desired to make estimates at the level of small areas within the region. Because of the frequency with which this example arises in practice, small-domain estimation is often referred to as *small-area estimation*, even when the domains are not defined geographically. In describing here the major modelling issues that arise in small-area estimation we follow closely the notation and spirit of Chambers (1997); other useful references on the subject include Ghosh & Rao (1994) and Draper *et al.* (1993).

Consider a continuous survey variable *y* defined on a population *U* with known overall size *N*. A sample *s* of size *n* units is drawn randomly from *U*, with the main target being the population total *t* or mean $\bar{y}$ of *y*. Let *r* stand for the unsampled units in the population, and let *U* be divided into small areas $a = 1, \ldots, A$, with known sizes N_a. After the sample is drawn one can divide it up into area-specific subsamples, of sizes n_a, and a secondary goal is the estimation of the area totals t_a or means $\bar{y}_a$. In many cases this cannot be done without the aid of a model that suggests how information should be combined across the areas, to improve the estimation within a given area (another name for this idea is *borrowing strength* from all the areas to estimate t_a and $\bar{y}_a$ for each area *a*).

Given a vector of *p* covariates **x** which are related to *y* and available on each unit in *U*, a typical model-based approach to small-area estimation would assume a linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathrm{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{V}(\mathbf{e}) = \sigma^2\mathbf{V}. \tag{9.22}$$

Here **y** is the vector of length *N* containing all the population values of *y*, **X** is the $N \times p$ matrix of population covariate values, **β** is a *p*-vector of regression coefficients, **e** is an unobservable vector of errors, and the covariance matrix **V** of **e** is assumed known (and often diagonal). Under (9.22) a model-unbiased estimate of the population mean $\bar{y}$ is

$$\hat{\bar{y}} = \frac{1}{N}\left(\sum_s y_j + \sum_r \mathbf{x}_j^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right), \tag{9.23}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^{\mathrm{T}}\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s^{\mathrm{T}}\mathbf{V}_s^{-1}\mathbf{y}_s$$

and the subscript *s* in the equation below (9.23) refers to the sub-vectors and sub-matrices consisting only of the sampled units. This is a typical prediction-style estimator of $\hat{\bar{y}}$ (see Chapter 2): the sampled values of *y* in the population are used directly in the estimation of the total, and the unsampled values of *y* are predicted by the model.

Probably the most widely used method for estimating the means $\bar{y}_a$ of the small areas in the context of a model such as the one above is *synthetic estimation*. The key assumption in this approach is that the same linear model (9.22) holds in each small area, that is, the relationship between *y* and *x* is constant across domains. Under this assumption it is sensible to estimate **β** from the entire sample, but then mimic the first line of (9.23) in each area separately:

$$\hat{\bar{y}}_a = \frac{1}{N_a}\left(\sum_{s_a} y_j + \sum_{r_a} \mathbf{x}_j^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right), \tag{9.24}$$

where s_a and r_a are the sampled and unsampled units in area a. However, the homogeneity assumption underlying (9.24) may well not be true. Making this assumption creates an estimate of $\bar{y}_a$ with relatively low variance (because the estimate of $\boldsymbol{\beta}$ borrows strength across the whole sample) but potentially large bias in any given area (if the constant-$\boldsymbol{\beta}$ assumption is far from true).

One way to avoid the area-level bias potentially inherent in (9.24) is to estimate the relationship between y and x separately in each domain, by fitting the model

$$\mathbf{y}_a = \mathbf{x}_a^{\mathrm{T}}\boldsymbol{\beta}_a + \mathbf{e}_a, \qquad \mathrm{E}(\mathbf{e}_a) = \mathbf{0}, \quad \mathrm{V}(\mathbf{e}_a) = \sigma_a^2\mathbf{V}_a. \tag{9.25}$$

The *direct estimate* of $\bar{y}_a$ suggested by this model is then

$$\hat{\bar{y}}_a = \frac{1}{N_a}\left(\sum_{s_a} y_j + \sum_{r_a} \mathbf{x}_j^{\mathrm{T}}\hat{\boldsymbol{\beta}}_a\right) \tag{9.26}$$

where

$$\hat{\boldsymbol{\beta}}_a = \left(\mathbf{X}_{s_a}^{\mathrm{T}}\mathbf{V}_{s_a}^{-1}\mathbf{X}_{s_a}\right)^{-1}\mathbf{X}_{s_a}^{\mathrm{T}}\mathbf{V}_{s_a}^{-1}\mathbf{y}_{s_a}$$

that is, simply estimate separate regressions in each area. This estimator is model-unbiased in each domain but will typically have large variance, since the domain-level sample sizes are usually small.

Thus each of the synthetic and direct estimates has potential flaws, in the directions of large bias and large variance, respectively, which suggests searching for a compromise estimator. The standard choice is based on an expansion of model (9.22),

$$y_j = \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta} + \sum_a I(j \in a)\alpha_a + e_j, \tag{9.27}$$

in which $I(p)$ is 1 if proposition p is true and 0 otherwise, and the (unobservable) α_a are referred to as *area effects* (model (9.22) just corresponds to the special case that all of the area effects are zero). If the α_a are regarded as IID random variables with mean 0 and variance σ_α^2, the resulting specification is a *random-effects* model; if the α_a are simply parameters (representing area-specific deviations from a common intercept) summing to 0 across the domains, the result is a *fixed-effects* model. Under either specification the compromise estimator takes the form

$$\hat{\bar{y}}_a = \frac{1}{N_a}\left[\sum_{s_a} y_j + \sum_{r_a}\left(\mathbf{x}_j^{\mathrm{T}}\hat{\boldsymbol{\beta}}_a + \hat{\alpha}_a\right)\right], \tag{9.28}$$

in which estimates of $\boldsymbol{\beta}_a$ and the α_a are obtainable from standard *multi-level modelling* software such as *ML*wi*N* (Goldstein *et al*. 1997). Choosing between fixed- and random-effects formulations depends on the number of small areas and the relative magnitude of the within- and between-area variation in outcomes of interest: for example, with a large number of domains and a fairly large degree of between-area homogeneity, a random-effects model would be indicated.

When the domains are in fact geographic areas and there is reason to believe that adjacent areas should exhibit similar responses, variations on (9.27) based on *spatial smoothing* are possible; see Cowling *et al*. (1996). Other refinements of the methods described here include *empirical Bayes* smoothing of direct estimates (see, for example, Draper *et al*. 1993) and small-area estimation of counts rather than continuous outcomes, based on *SPREE* estimates (for example, Purcell & Kish 1980).

As an example of these ideas in action, the UK Office for National Statistics (ONS) has in the past used a version of direct estimation in the Annual Business Inquiry (ABI; see Section 9.5 for analysis of some ABI data). The basic idea was that a ratio-type estimator (based on regression through the origin) was fitted for each survey variable with a different parameter in each stratum, and then – based on auxiliary data from the register – a complete "survey" record was made for each non-sampled business in the population to supplement the sample responses (this is an application of equation (9.26)). This allowed cross-tabulation of results for very small domains in more or less any conceivable combination, but did not make any comment about the quality of the data. In effect an estimate for a region (not part of the survey stratification) was made up of an estimate of each cell in the region by stratum cross-tabulation, with the appropriate direct estimators added to give an overall estimate. This relies heavily on the assumption that the strata define all the variability in the data, and that the samples are representative.

In current practice at ONS, most small-area estimates are for domains which coincide with strata, and hence are not subject to the uncertainty arising from having to estimate the domain size. Two surveys (one extant and monthly, the other planned and annual) make domain-type estimates for regions from data which are not stratified by region, and then constrain these estimated totals (a) to reproduce known auxiliary variable totals and (b) to sum to the same overall estimate for the UK (a kind of benchmarking; see section 9.3). Some ONS surveys (normally in which the sample size per stratum is small) use combined ratio estimation, which is based on the assumption of a constant ratio (or regression slope) over the size strata; this is similar to the synthetic estimation method (9.24) above. ONS does not at present use multi-level models, of the type leading to estimators such as (9.28), in business surveys, but plans to explore their use in the future.

The effects of model assumption errors similar to those in small-area estimation will be explored in Section 9.8.

9.7 Non-ignorable nonresponse

In chapter 8 we discussed the effects of nonresponse errors on business surveys. Three types of data missingness at the unit level were defined in that chapter: letting R_k be 1 if sample unit k responds and 0 if not,

- *missingness completely at random* occurs when R_k is stochastically independent of the outcome(s);
- *missingness at random given an auxiliary variable* X_k occurs if R_k is conditionally independent of y_k given x_k; and
- *informative* or *non-ignorable* missingness occurs when R_k and y_k remain dependent even after conditioning on (adjusting for) all available auxiliary variables x_k.

The two versions of missingness at random in this list are referred to as *ignorable nonresponse*, because in those cases no bias in estimating aspects of the population distribution of y is induced by the missingness (although, as chapter 8 points out, missingness at random will inflate observed sampling variances, because the obtained sample size is smaller than the planned sample size). The main problem created by *non-ignorable nonresponse* (NINR) is that, when it occurs, estimates based only on the respondents will be biased. In the setting of stratified random sampling examined in section 8.3, for example, equation (1) of that section summarized the bias in estimating the population total t of y as

$$\text{bias}(\hat{t}_r) = \sum_{h=1}^{H} N_h (1 - \overline{R}_h)(\mu_{h,R=1} - \mu_{h,R=0}), \qquad (9.29)$$

where $\hat{t}_r = \sum_{h=1}^{H} N_h \bar{y}_{rh}$ is the estimate of t based only on the responding units and in which $\mu_{h,R=1}$ and $\mu_{h,R=0}$ are the means in stratum h for the respondents and nonrespondents, respectively. NINR implies that these two means are not equal, and the greater the disparity between them, the larger the overall bias in (9.29). Thus with NINR it is not necessarily adequate simply to act as if the respondent data set, of total sample size $n_r = \sum_{h=1}^{H} n_{rh}$, is equivalent to what one would have obtained with an intended sample of size n_r that had no missingness.

There is an operational problem with this conclusion, though: how can one judge whether the missingness is ignorable, when by definition the y_k values are not observed for the units with $R_k = 0$? One approach to answering this question in longitudinal surveys is to consult the frame for variables that are good proxies for y, for example, y in period t may be strongly correlated with y in period (t–1), and y_{t-1} may well be available for many of the units for which R = 0 at time t; or one may be able to compare sample respondents and nonrespondents at time t with respect to their values on auxiliary variables x, which have in the past been strongly correlated with y, at time (t–1).

An even greater difficulty is what to do about NINR when it is suspected. Assuming NINR, the only way forward is evidently through the specification of a model which predicts what the observed y values would have been for the units for which $R_k = 0$. There appears to have been little or no systematic attempt in the literature to tailor the construction of such models to the business statistics framework (for example, the UK Office for National Statistics makes no use of NINR models in the analysis of any of its business survey results at present). Attempts have been made in other settings, however, and in the rest of this section we review two leading methods that appear of potential relevance to business statistics.

9.7.1 Selection models for continuous outcomes

Copas & Li (1997) analyse data from a local skills audit conducted as a sample survey in Coventry, UK, in 1988. In one analysis of n = 1435 adults known to be in full-time employment (and assumed to be randomly sampled from the population of such adults in Coventry), the outcome of interest y was income (pounds per week), with gender and age as the principal auxiliary (x) variables. There was no missingness on the x-values, but 8% of the adults refused to provide income information, yielding a complete-cases sample size of 1323. A response rate of 92% may seem admirably high, but there was good reason to believe that the probability of nonresponse was a function of income.

Copas & Li used *selection models*, an approach dating back to the 1970s in the econometric literature (see, for example, Heckman 1979), to quantify the possible effects of NINR in this problem. Along with the observed y and $\mathbf{x}$ values, where $\mathbf{x}$ is in general a vector, the basic idea of these models is to posit the existence of an unobserved, or *latent*, variable z which represents the propensity to respond in the survey, and to relate $(y, \mathbf{x}, z)$ by the pair of regression equations

$$\begin{aligned} y_k &= \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\beta} + \sigma e_k \\ z_k &= \mathbf{x}_k^{T} \boldsymbol{\gamma} + \varepsilon_k \end{aligned}, \qquad (9.30)$$

in which the pair (e_k, ε_k) is taken to be bivariate normal with $\mathrm{E}(e_k) = \mathrm{E}(\varepsilon_k) = 0$, $\mathrm{V}(e_k) = \mathrm{V}(\varepsilon_k) = 1$, and $\mathrm{corr}(e_k, \varepsilon_k) = \rho$. The first equation in (9.30) might be termed the *observation* equation, the second the *selection* equation, and application to missing data in surveys arises by assuming that y is only observed if the latent variable z is positive. The correlation between the error terms in the two equations captures the premise that (i) e_k is a kind of place-holder for a set of unobserved auxiliary variables $\mathbf{x}_y$ that would help to predict y if they had been observed, (ii) ε_k similarly "contains" another set of unobserved auxiliary variables $\mathbf{x}_z$ that would help to explain the propensity to respond if they had been measured, and (iii) the two sets of variables in $\mathbf{x}_y$ and $\mathbf{x}_z$ are likely to overlap, inducing a correlation between e_k and ε_k. If $\rho = 0$ there is no information in the selection equation for predicting y, which implies ignorable nonresponse, but if $\rho \neq 0$ then y is subject to NINR.

Copas & Li fit model (9.30) by profile maximum likelihood (see Draper & Cheal 1998 for a Bayesian analysis) and examine the sensitivity of results to the possibility of NINR by calculating estimates of the population mean for y, and standard errors of those estimates, as a function of ρ. They note that "For a well-designed and well-executed survey such as [the Coventry skills audit] it is implausible that $|\rho|$ would be very large. With an overall [nonresponse] rate of 8%, a fairly extreme possibility might be that the probability of missing data at the lower quartile of [the distribution of] y is 4% whereas at the upper quartile it is 12% (three times as large)." This gives a plausible range for ρ between –0.40 and 0.40, leading to bias-adjusted population mean estimates in the range (138, 148) pounds per week as compared with the unadjusted estimate $\bar{y}$ of 142. Thus with a nonresponse rate of only 8%, the bias correction to adjust for NINR in this example is only about 3–4% of the unadjusted estimate, but this is of the same order of magnitude as the standard error of y, so that (since it is not clear whether the bias is positive or negative) "the extra uncertainty [attached to $\bar{y}$ arising from the possibility of NINR] could be thought of as doubling the [variance] of estimation."

This provides a concrete summary of the possible effects of NINR and (ideally) what to do about these effects: when unit-level nonresponse occurs in a survey, if both the direction and the magnitude of biases introduced by the nonresponse can be quantified, based on reasonable modelling and past experience, then bias adjustment should be undertaken; and if the direction and magnitude are hard to pin down, then the standard uncertainty bands based only on the observed data should widen to acknowledge the possibility of non-ignorable nonresponse.

9.7.2 Pattern-mixture models for categorical outcomes

Forster & Smith (1998) examine data from the 1992 British general election panel survey to quantify the effects of possible NINR on estimates of voting intention y (which was categorical at four levels). In their random sample of 1242 individuals the available auxiliary variables were gender and social class (categorical at five levels), which were known for all sampled people, but 375 (30%) of the sampled individuals refused to make their voting intention known. Denoting the vector of auxiliary variables by $\mathbf{x}$ and the response indicator by R, the problem (as above) is to construct joint probability models for $(y, \mathbf{x}, R)$ that will permit imputation of what the observed voting intentions would have been for those people for whom $R = 0$. Maximum likelihood estimation of voting intent based solely on the observed y-values in the survey yielded (Conservative, Labour, Liberal Democrat, Other) = (C, L, LD, O) = (45.6%, 34.3%, 17.2%, 3.0%).

Using the notation of conditional independence developed by Dawid (1979), the assumption of missingness completely at random corresponds to $R \perp \{\mathbf{x}, y\}$ (that is, R is independent of $(\mathbf{x}, y)$), whereas missingness at random given $\mathbf{x}$ is expressed as $R \perp y \mid \mathbf{x}$. All other models assume NINR in one form or another. Different modelling strategies correspond to different factorisations of the joint distribution $p(y, R, \mathbf{x})$, for example, the factorisation

$p(y,R,\mathbf{x}) = p(\mathbf{x})p(R \mid \mathbf{x})p(y \mid \mathbf{x},R)$ reduces, under the assumption of missingness at random, to the model

$$p(y,R,\mathbf{x}) = p(\mathbf{x})p(R \mid \mathbf{x})p(y \mid \mathbf{x}) \tag{9.31}$$

for the fully observed data, which is in the class of *decomposable graphical log-linear models* (see for example, Dawid & Lauritzen 1993). This model, approached in a Bayesian way but with prior distributions on the parameters with little information content, yielded with the above survey results – as it must – results in close agreement with the maximum likelihood estimates: (C, L, LD, O) = (44.8%, 35.0%, 17.1%, 3.1%), with 95% uncertainty bands [(41.3, 48.3), (31.6, 38.5), (14.5, 19.7), (2.0, 4.5)].

In their central NINR modelling Forster & Smith employ the factorisation

$$p(y,R,\mathbf{x}) = p(R,\mathbf{x})p(y \mid R,\mathbf{x}), \tag{9.32}$$

a *pattern-mixture* specification (for example, Glynn *et al.* 1986). Forster & Smith's main approach is as follows:

> "As we are only considering non-response on *y*, $n(R,\mathbf{x})$ [the cross-tabulation of *R* against **x**] and $n(y,\mathbf{x},R=1)$ are fully observed. Hence, we have all the information required for inference about $p(R,\mathbf{x})$ and $p(y,\mathbf{x} \mid R=1)$. However, *y* is completely missing when *R* = 0 and so ... any inference for $p(y,\mathbf{x})$ requires some kind of prior information concerning $p(y,\mathbf{x} \mid R=0)$. This prior distribution ought perhaps to be referred to as the *subjective distribution*, as it remains unaltered in the light of the observed data. ... An intuitively attractive and computationally straightforward approach is to consider the parameters $p(R,\mathbf{x})$, $p(y \mid \mathbf{x}=x',R=1)$ and $\theta_{x'}$, $x'=1,\ldots,t$ [where *t* is the number of distinct values taken by **x**]. The parameters $\theta_{x'}$ represent the extent of prior belief in non-ignorability. If $\theta_{x'}=0$ then this corresponds to ignorability of nonresponse for stratum x', and if all $\theta_{x'}=0$ then $y \perp R \mid \mathbf{x}$ and non-response is [missing at random given **x**]. ... Hence, the $\theta_{x'}$ are easy to interpret and prior information regarding ignorability may be straightforwardly incorporated into the model via a prior distribution. ... We choose to use multivariate normal distributions for $\theta_{x'}$, with mean $\mu_{x'}$ and variance $\sigma^2_{x'}$ determined by the prior belief concerning the extent and structure of non-ignorability."

The parameters $\theta_{x'}$ in this formulation play the role of the correlation parameter ρ in the Copas & Li approach in section 9.7.1.

There was evidence from the literature that nonrespondents to polls in British general elections prior to 1992 were more heavily pro-Conservative than respondents. Using a reasonable prior specification based on this evidence, Forster & Smith obtained adjusted estimates of (C, L, LD, O) = (47.6, 33.0, 16.5, 2.9), with 95% intervals [(42.1, 53.0), (28.7, 37.6), (13.6, 19.7), (1.9, 4.2)]. In comparison with the results above based only on the respondents, the bias adjustments were on the order of 2-3 percentage points for the two largest political parties, increasing the estimated lead of the Conservatives over Labour by 5 percentage points (a large difference in practical terms), and the 95% uncertainty bands were on average 34% wider after the possibility of NINR was accounted for.

Forster & Smith also provide a useful formula for sample size calculations at design time to anticipate possible NINR: in their framework, "the effect of allowing for non-ignorability is to reduce the effective observed sample size $n(R=1, X=x)$ in stratum x to

$$\frac{n(R=1, \mathbf{x}=x')}{1+\frac{n(R=1, \mathbf{x}=x')\sigma_x^2\left[n(R=0, \mathbf{x}=x')/n(\mathbf{x}=x')\right]^2}{s}}, \tag{9.33}$$

[where s is the number of observed levels of y]. The proportions of respondents and nonrespondents in each stratum will not be known in advance and a prior estimate will be required," as will a prior specification of $\sigma_{x'}$, the amount of uncertainty about how strong the NINR will be in stratum x'. These things may not be easy to specify at design time, but that is typical of survey design, and in any case (9.33) can serve as the basis of a sensitivity analysis.

Effects of errors in modelling assumptions similar to those arising from attempts to cope with non-ignorable nonresponse will be considered in the next and final section of this chapter.

9.8 Conclusions

We have seen in the previous sections that models are ubiquitous in the analysis of business surveys. Since a statistical model is nothing more (or less) than a collection of assumptions about the relationship between observed and unobserved data, and since by their nature some of these assumptions are not known to be valid with certainty, assessing the impact of errors in modelling assumptions is evidently crucial to the success of business surveys that employ them. Three examples of this arising from Sections 9.3, 9.6 and 9.7 are as follows.

- On the topic of models for benchmarking, Cholette & Dagum (1994) admit that "In real cases, the gain in efficiency from the regression method [for benchmarking which they advocate] will depend on how well the *ARMA* models [for the monthly series to be benchmarked] are identified and estimated."
- In small-area estimation Chambers (1997) concludes that "At the time of writing a general consensus on an appropriate 'robust' methodology for measuring the 'overall reliability' of small-area estimates has not been reached," which is one way of saying that model assumption errors in small-area estimation may well dominate other sources of error.
- With regard to non-ignorable nonresponse, Forster & Smith (1998) report on the results of a follow-up survey of the 1242 original participants in the 1992 British general election panel survey: "21 individuals did not respond and 86 claimed not to have voted. Of the remaining 1135, 44.1% [reported voting] Conservative, 32.2% Labour, 21.0% Liberal Democrat, and 2.8% other. Of these, 317 were nonrespondents to the original survey, for

whom the corresponding proportions were 41.0%, 25.6%, 30.0%, and 3.5%." Thus in the end the original nonrespondents reported voting in a way that was wholly unanticipated – far more strongly for the Liberal Democrats (LDs) than any experts would have predicted – yielding an overall percentage for the LDs that fell outside the 95% interval from the pattern mixture modelling (even with its much wider uncertainty bands). This highlights the fact that even when reasonable modelling assumptions are employed based on expert knowledge, occurrences outside the realm of plausible prior expectation can be left unanticipated by the modelling.

It would appear that best practice in dealing with model assumption errors in business statistics matches the situation in statistical modelling quite generally, in that two main tools are available:

- The sensible use of *model diagnostics* (see, for example, Cook & Weisberg 1982); and
- A willingness to employ *sensitivity analysis* (see, for example, Skene *et al*. 1986): varying the modelling assumptions across plausible ranges to discover their effects on the estimates of the quantities of principal interest. This will often involve *simulation studies* (see, for example, Hammersley & Handscomb 1979). In the class of linear models, for example, a suggestive (but not exhaustive) list of categories of modelling assumptions worth exploring might include the following:
    - *Transformation* of outcome *y* and one or more predictors *x*;
    - Choice of the *functional form* by which *y* and the *x*'s are related;
    - Assumptions about the *variance structure* and *distribution* of the error terms in the model;
    - *Choice of predictor variables* from among a potentially large set of *x*'s; and
    - *Choice of outlier treatment* method.

Both of these approaches, including a number of the model assumption categories listed here, were illustrated in Section 9.5 on cut-off sampling. Figure 9.1 gives a scatterplot of returned turnover against register employment in a simulated population based on the 1995 UK *ABI* survey, and a residual plot obtained from fitting a regression through the origin with both variables on the raw scale. Both plots show (a) a number of *high-leverage* points (Weisberg 1985) – companies exerting a large influence on the estimated slope, which can dramatically shift the ratio estimator based on the regression model (9.17) – and (b) a strong bunching up of points near the origin, which implies that the weighted least squares method used to estimate the slope may not be making the most efficient use of the data.

Each of these problems suggests alternative modelling assumptions. Difficulty (a) is a *robustness* problem (Huber 1981), perhaps most simply solved by means of *trimmed regression*: set aside a small proportion of the companies with the highest register employment, and fit model (9.17) to the remaining data. Difficulty (b) suggests a *data-analytic* solution (see, for example, Mosteller & Tukey 1977) based on *variable transformation*: instead of regressing *y* on *x*, regress $\log(y)$ on $\log(1+x)$. This line of reasoning yields three main cut-off estimators, based on three different models: (i) regression through the origin on the raw scale, employing all of the data; (ii) regression through the origin on the raw scale, trimming the high-leverage companies; and (iii) ordinary least-squares regression using all of the data on the log-log scale.

Evaluating the quality of these estimators is an exercise in sensitivity analysis based on simulation: one may (A) repeatedly generate simulated populations similar to the reality in which the chosen cut-off estimator will be employed, computing the true population total turnover in each simulation repetition; (B) compute each of the three cut-off estimates for each simulated population; and (C) evaluate the estimation methods in terms of such summaries as relative bias and/or root mean squared error. The results, in Table 9.1, show clearly that – for populations like the *ABI* data – trimmed ratio estimation on the raw scale and regression estimation on the log-log scale perform well. This does not prove that these methods would work equally well on other populations; simulation-based sensitivity analysis of this type must be employed on a wide variety of population types to draw such a conclusion, and an interaction between population type and estimation method may well be found: method (ii) works best with population type I, method (iii) works best with type II, and so on.

There is another variety of sensitivity analysis worth mentioning as well: examining the effects of model assumptions *on a single* (*real*) *sample* rather than across a number of simulated populations and samples. In this approach one makes a list $\{A_1,\ldots,A_k\}$ of modelling assumptions that all seem to be plausible for the given sample, based on expert judgement and model diagnostics, and then one computes the corresponding conclusions $\{C_1,\ldots,C_k\}$ resulting from the set of assumptions. The results of this type of sensitivity analysis may be summarised either qualitatively or quantitatively, as follows.

- *Qualitative summary*. The idea is simply to see if "all reasonable roads lead to Rome," that is, to see if across the span of plausible $\{A_1,\ldots,A_k\}$ the resulting $\{C_1,\ldots,C_k\}$ largely agree with regard to the quantities of principal interest. If they do, then confidence increases that model assumption errors do not play a large part in the threats to the survey's validity. If they do not, then this approach is more problematic; one is left with a qualitative summary of the form

$$\text{If assumptions } A_1 \text{ then conclusions } C_1, \text{if } A_2 \text{ then } C_2,\ldots \qquad (9.34)$$

  which may well not be satisfactory as a basis for decision-making based on the survey.
- *Quantitative summary*. To go beyond (9.34) one must be willing to place weights on the relative plausibility (that is, probabilities) of the assumptions A_i, to produce a composite summary that reflects both within-model and between-model uncertainty. There is now a well-developed Bayesian approach to doing this (for example, Draper 1995): with *y* as an outcome to be predicted, model ξ_i (based on assumptions A_i) given probability p_i and leading to predictive distribution for *y* with mean and standard deviation (SD) $\hat{\mu}_i$ and $\hat{\sigma}_i$, respectively,

$$\begin{aligned}\hat{\mathrm{V}}(y) &= \mathrm{V}_\xi\left[\hat{\mathrm{E}}(y \mid \xi)\right] &+\ & \hat{\mathrm{E}}_\xi\left[\hat{\mathrm{V}}(y \mid \xi)\right] \ = \hat{\sigma}^2 \\ &= \sum_{i=1}^{k} p_i(\hat{\mu}_i - \hat{\mu})^2 &+\ & \sum_{i=1}^{k} p_i \hat{\sigma}_i^2 \\ &= \begin{array}{c}\text{between - model} \\ \text{variance}\end{array} &+\ & \begin{array}{c}\text{within - model} \\ \text{variance}\end{array}\end{aligned} \qquad (9.35)$$

where

$$\hat{\mathrm{E}}(y) = \mathrm{E}_\xi\left[\hat{\mathrm{E}}(y \mid \xi)\right] = \sum_{i=1}^{k} p_i \hat{\mu}_i = \hat{\mu}\,. \qquad (9.36)$$

Thus the overall predictive uncertainty about *y* decomposes into the sum of {the uncertainty conditional on a given set of modelling assumptions} and {the uncertainty about the modelling assumptions themselves}. There may be substantive and technical difficulties in implementing this approach in practice, however, and it has not yet been attempted with business survey data; this type of *model uncertainty audit* is in the category of possible future best practice in business surveys.

We conclude this section, and the chapter, by summarizing the above discussion.

Recommendation: Best-practice reporting in business surveys involving *model-based methods* should

- Use a blend of model diagnostics, simulation studies, and qualitative sensitivity analyses to make consumers of the survey aware of (a) the plausibility of the principal assumptions made by the models employed and (b) the effects of varying these assumptions, across reasonable alternative specifications, on the summary estimates of principal interest.

Part 3: Other Aspects of Quality

10 Comparability and coherence

Eva Elvers[5], Statistics Sweden

10.1 Introduction

Coherence relates to sets of statistics and takes into account how well the statistics can be used together. Statistics are estimates of finite population parameters (FPPs), as described in previous chapters and in the next section. The target is rarely achieved for many reasons. The smaller the discrepancy between the value of a statistic and its target, the more accurate is the statistic.

A statistic can be considered as consisting of the sum of the FPP and an estimation error. There are two principal error parts, systematic errors (that may lead to a bias) and random errors. The producer normally aims at the bias being nil or negligible, and also at random errors being small (close to zero in absolute or relative terms). One way of describing the inaccuracy is through the root mean square error, another is an uncertainty interval. The interval could be symmetric around the point-estimate.

The user has a set of FPPs in mind that he/she wishes to study. Then there may be statistics published that suit these wishes – "off-the-shelf" – but often it is necessary to use several sets of statistics. Such a usage may include combination of several FPPs into new ones. The user needs to know if there are statistics with target FPPs that are equal – or at least close – to his/her "ideal".

Coherence is a more general concept than comparability. Questions on coherence arise for example when production statistics and foreign trade statistics are used together, or production statistics and employment statistics, or annual statistics and short-term statistics.

In quality reports to Eurostat, comparability and coherence are two quality components. Since these components have much in common – the former being a special case of the latter – they are here described and discussed in a single chapter. Obviously, comparability between Member States (MSs) is important to Eurostat, and also comparability between countries in general. Comparability over time is another comparability aspect. At present Eurostat does usually not include comparisons between non-geographical domains in the comparability component.

Coherence aspects are discussed below first with emphasis on the user in Section 10.2 and then with emphasis on the producer in Section 10.3. The structure is largely the same in both cases, using six sub-headings, mainly as below

1. definitions in theory
2. definitions in practice
3. accuracy and consistent estimates

[5] Several persons have contributed with comments and examples, especially Ole Black and Mark Williams at ONS

4. comparability over time
5. international comparability
6. concluding comments

The examples all refer to business statistics, but the theory is general for official statistics. Section 10.4 is more illustrative, based on some national situations. Summaries and conclusions are given along with the text, largely in Sections 10.2.6 and 10.3.6.

10.2 Coherence – emphasising the user perspective

10.2.1 Definitions in theory

As stated previously, statistics are estimates of finite population parameters (FPPs). Ingredients in such a parameter are

- statistical measure (total, mean, median, etc);
- variable (production, number of hours worked, etc);
- unit (enterprise, kind-of-activity-unit, etc);
- domain (sub-population, for example defined by a standard classification like NACE Rev. 1);
- reference times; both units and variable values relate to specific times.

The reference times are mostly time intervals, like a calendar year, a quarter, or a month. (However, some variables may refer to a point in time, for example the starting point of the period.) Usually reference times agree for all variables and units in a FPP. This means for example for monthly statistics that the delineation of units should refer to the current month. It follows from the above that units, classifications, other auxiliary variables, and reference times are essential to consider whenever using statistics.

In a joint use of several sets of statistics, the user wishes to keep some of the ingredients of the FPPs constant and vary one or more of the others. Some typical examples, with emphasis on what is varied:

– comparison over time: reference times, for example every month from a given one onwards;
– comparison of countries: domains are Member States or other countries;
– comparison between non-geographical groups: domains like industries are varied;
– new statistics using several surveys: combining statistics from different business surveys (production & employment, annual & short-term) for further analysis of industries for example.

A simple example of a complex setting is: first taking ratios between production and number of hours worked using two surveys and then comparing those relative quantities over different aspects of space: geographical areas, industries, size groups etc. To this end, the surveys should be equal in their units, domains, and reference times. The domains are defined by for example an industrial classification that needs to be the same for all surveys.

When a user is judging coherence, definitions of the target finite population parameters (regarding units, population and domain delineation, variables, and reference times) play a primary role. Accuracy is important, but it plays a secondary and different role. The more accurate the statistics, the smaller the disturbances; the study is more easily performed, and the conclusions drawn are usually stronger.

10.2.2 Definitions in practice

As described in the previous section, joint use of sets of statistics builds on some ingredients of the target statistics being the same. The difficulties meeting the user often depend strongly on the "distance" between the statistics used jointly. It may not be trivial even within a single survey, since definitions can vary (for example for production and employment, reference times could be a period for one and a point in time for the other). Normally, however, the problems increase considerably when using several surveys.

Even if definitions are the same in principle – as far as the user can see – they may differ in practice. One survey may have the reference time of the domains equal to that of the variable and the other use that of the frame (which the quality reports should show). A further example is the enterprise unit; it has to be defined and applied in the same way in both surveys. In a comparison between MSs, the enterprise definition may vary a lot, in spite of there being a Regulation on statistical units.

In practice there is an influence from the methodology used for example in data collection and estimation. Hence, the user needs information also on such influential factors.

10.2.3 Accuracy and consistent estimates

Accuracy has, of course, to be considered when studying for example how the ratio between production and hours worked varies over industries, so that differences that can be due only to "noise" are not stated to be significant. The user needs a measure of the overall accuracy in the joint use. This means an assessment of inaccuracy from all sources, not only due to taking a sample. It is important that the measure is realistic.

If there is a relationship between the FPPs involved, many users find it convenient if the estimates also fulfil this relationship. Two simple examples:

(i) The number of employees in two different surveys (on employment and production) with definitions such that the FPPs are equal.

(ii) Monthly and annual production statistics with definitions such that the sum over the twelve calendar months equals the annual value.

The expression *consistent estimates* will be used here to emphasise that the estimation procedures have forced the estimates to have the same relationship as the FPPs, see Section 10.3.3 for some detail. Obviously, statistics can be coherent without giving consistent estimates. This is normally the case with preliminary and definitive statistics. *Note* that the concept of consistent estimates is different from consistency in asymptotic theory.

If a user has two statistics that he/she believes estimate the same FPP and these estimates differ more than expected, from the inaccuracy measures given, the user should suspect deficiencies in coherence. A simple example is as follows. Without going into technical details, assume that uncertainty intervals are given.

1) The figures are 750 ± 25 and 705 ± 10
   These are not coherent from what can be seen.
   This signals that there are differences in definitions that have not been stated or the user has not observed. Another possibility is that one or two of the intervals is too short.

2) The figures are 700 ± 25 and 705 ± 10
   These are coherent from what can be seen.
   It would be more convenient for the user to have a single figure (consistent estimates), say 704 ± 9

The discussion in this section has emphasised the random part of the estimation error. There may also be systematic errors to take into account when using statistics. Such errors could be caused for example by the data collection. The distinction between definitions and systematic deviations is not always clear-cut, though, since definitions in practice are influenced by many factors in, for example, data collection and estimation.

10.2.4 Comparability over time

Comparisons over time are frequent. There are often two conflicting user interests as to the statistics to be produced:

- stability of definitions to compare the present with the previous for a special issue;
- the current state should be well described.

The first one works in the direction of comparability, whereas the second one goes in the opposite direction. This may be a cause of tension in statistical systems. When a change is made, special actions are often taken to improve the comparability, for example by producing statistics in both ways on one occasion or even re-estimating a part of the old series in terms of the new definitions.

There may be different opinions as to whether it is more important to estimate the level or the change accurately – different statistics may have different priorities. Short-term statistics often emphasise changes. To make that possible, comparability is needed over the time period that the changes refer to. Users of annual statistics may find the level to be more important. The National Accounts need to describe both level and change.

A further aspect of comparability over time is that certain users (for example using economic statistics indicating short term changes) are anxious to be able to separate for example

- ♦ trend and
- ♦ regular seasonal variations.

Technical means for this purpose are seasonal adjustments and calendar adjustments. To include such parameters is an enrichment of the statistics.

10.2.5 International comparability

A particular, important aspect is comparability between Member States, other countries and geographical areas in general. This involves not only different producers of the statistics but also further differences due to inherent dissimilarities between countries: labour market rules, economic practices, tax rules, etc.

Attempts to reduce differences – to increase comparability for the benefit of the user – by using similar concepts and definitions have been going on internationally for a long time; they are time-consuming tasks. There are many activities for harmonisation in business statistics in Eurostat and other international authorities, see Section 10.3.

10.2.6 Some user-based conclusions

In summary, comparability and coherence within and between sets of statistics require some definitions to be the same, for example units, variables, or reference times, depending on the particular joint use. The user needs information on differences and their consequences from the producer. The quality report for a certain set of statistics should provide such information with regard to comparability over time and coherence with other sets of statistics. It is not possible to include all other sets but experience should be used to list uses that are frequent and where users are likely to need help.

Comparability and coherence in general depend on definitions. Accuracy plays a different role. There is, however, not always a clear-cut distinction. Definitions may seem clear and unambiguous in theory but still vary in practical work. There may be a tendency not to include such deviations when measuring accuracy, although that should be done. If, for example, there is an undeclared systematic deviation in one survey but not another, there will be deficiencies in coherence between the two sets of statistics.

As a consequence of the above, comparability and coherence depend on the "distance" between producers; the deficiencies mostly increase in the following order: parts of a single survey, different surveys at the same agency, different organisations in the same country, statistical offices in different countries.

It is important for the user to have accuracy measures when using statistics together. It is convenient if the joint use has been foreseen and prepared, for example so that estimates are consistent. Explanatory comments in cases of differences are helpful, for instance when there are substantial revisions.

10.3 Producer aspects on coherence, including comparability

10.3.1 Definitions in theory

The means of the producer to achieve coherence are several. To use the same definitions is, of course, one of them – to be consistent within the authority and with international standards.

There are many harmonisation activities internationally, and different activities have gone on for a long time in different fora. There is, for example, much effort at the EU level, performed by Eurostat and other authorities.

There are statistical standard classifications, like NACE for activities. There are also classifications for products. Furthermore, there are regulations on Business Registers (BRs), and on statistics, like Structural Business Statistics (SBS) and Short-Term Statistics (STS). There is a Regulation on statistical units for the observation and analysis of the production system in the Community. Unit delineation and the BR together form an important part of the basis of the statistics. The National Accounts are "at the top", building on a lot of other statistics and being one reason for coherence among them.

Even if there is a considerable set of definitions that have been agreed upon, this does not mean that there is full harmonisation. Interpretations and practices may still differ between countries.

10.3.2 Definitions in practice

There are many aspects to consider in the definition of a variable, both to achieve coherence between surveys and with international guidelines, and to make the measurement and data collection procedures easy and accurate. Respondents mostly provide information from their accounting systems, which advocates a choice of definitions in agreement with accounting systems in general use. Business organisation has to be considered carefully when defining and delineating both units and variables. An example here is how to handle production by bought-in employment.

Ideally there should be co-ordination activities between statistical surveys, for example in questionnaires, instructions to respondents, and data editing. This may be more straightforward within a National Statistical Institute (NSI) than between organisations.

The activities within an NSI may include the basics: units, delineation of population and domains, variables, statistical measures, and reference times, and also procedures like data collection and estimation. Using the same BR as a frame, constructing the frame at the same time, updating the units in the same way at the same time (with regard to business structure, classifications etc), addressing questionnaires to the same unit, etc are further actions influencing coherence and accuracy.

There may also be activities between organisations and different countries. Foreign trade statistics is a clear-cut example where investigations are possible through so-called mirror statistics; the exports of country A to country B should equal the imports of country B from country A. There are differences to be studied, largely due to inaccuracy, for example measurement errors, but also due to differences in definitions between countries.

Overall, there are several principles which can be used to achieve comparability and coherence in general, both within and between nations, more or less far-reaching. The European Statistical System goes for the subsidiarity approach, where each Member State may implement surveys in its own way, together with quality assessments. This is preferred to attempts to harmonise production and to documentation of differences, often leading to tables with lots of footnotes.

10.3.3 Accuracy and consistent estimates

As stated in Section 10.2.3, it is important for the user to have appropriate accuracy measures in his/her joint use of statistics. The accuracy may be difficult to quantify. There could for instance be measurement errors for units sampled with probability one (that do not contribute to the sampling errors). If such a unit has different respondents in different surveys, and one of the respondents only includes one of several branches, this is a measurement error with severe consequences, if undetected. The ratio between production and hours worked by industry may be affected if the missing part is large, and there is clearly a risk of the accuracy measure not including this error fully. Hence, there may be a false conclusion. It may be regarded as due to a non-sampling error; it could also be viewed as an underestimation of inaccuracy.

The example may seem exaggerated, but such things happen. The following overall, and vague, statement seems reasonable (and in line with the previous sections): the further apart the surveys are, the greater the risk of differences between them – differences that affect the accuracy, often in a way that is not easy to assess. The joint use of statistics with inaccuracy of different character is more difficult than to use statistics from the same survey where the errors are "related", perhaps because there are systematic deviations that cancel at least partly in comparative studies or because the random errors are correlated.

In line with the above, including the example in Section 10.2.3, consider statistics as consisting of the target parameter and an estimation error, and assume the simple case with a symmetric uncertainty interval around the point estimate. The shorter the length of the interval, the more accurate the statistics, and the stronger the statistical inference in the joint use of statistics, for example comparisons. As just discussed above, there is a risk of producing too short an interval, not taking all the error sources into account. The coherence concept is tied to the target. In joint use of statistics, certain parts of the targets involved need to be equal, as Section 10.2.1 illustrates.

Consider the ratio between total production and total number of hours worked with both statistics based on a sample survey. If they emanate from the same survey, they are based on observations on the same set of units. So, if a sample happens to contain mostly small units, this is so for both numerator and denominator. The ratio does not vary so much around the population value as it would with two different samples. A smaller variation is obtained not only with respect to the sampling error, but it can be expected to hold for several further error sources, for example measurement errors.

Hence, comparability and coherence aspects in general make it desirable to co-ordinate the production of statistics that are used jointly. The estimation procedure may be co-ordinated between surveys. This can be done at different stages, with different strengths, and with different aims.

The aim could be to give the user as simple and coherent a message as possible, that is to have a high degree of co-ordination of the output from different surveys. This is different from handling each survey on its own and from using auxiliary information with the single aim of improving accuracy.

10.3.3.1 Some comments on methodology, especially benchmarking

One method of co-ordinating statistical output is so-called benchmarking, where one set of estimates is forced to agree with another. This is a special case of consistent estimates introduced in Section 10.2.3. Typically, short-term statistics could be benchmarked on annual statistics, if the former (after aggregation to the calendar year) are an indicator of the latter. One reason could be to simplify for the user by unifying the two time series (ensuring that the monthly series has the same annual sum as the annual series), another to improve the accuracy of the short-term statistics. For this to be meaningful, the two sets of statistics should have the same target parameters for the calendar year.

The use of procedures to make estimates consistent may influence not only one but both sets of statistics. The implementation of benchmarking of, say, short-term statistics on annual statistics, involves comparisons. These may consider not only the macro level, but also the micro level. The evaluations performed may imply further edits for both short-term and annual statistics.

In cases like benchmarking short-term statistics on annual statistics, the former have been published when the latter appear. That means a revision, may be one or two years after the first publication (longer for January than for December), or even more. Many users will react badly to revisions in their time series. Advantages and disadvantages have to be balanced.

There are several methods for benchmarking, based on different approaches to the two time series as to what is fixed and what is random variation, see for example Cholette & Dagum (1994) with emphasis on survey errors, Durbin & Quenneville (1997) with emphasis on stochastic time series models (and also references therein), and the very recent Dagum, Cholette & Chen (1998).

There is a recent suggestion on co-ordination at the estimation stage by Renssen & Nieuwenbroek (1997), who call their procedure aligning estimates. Surveys with variables in common – variables that are observed in these surveys and have unknown population totals – are pooled and the common variables are used as regressors (in addition to variables with known totals). Then the estimate obtained is used as auxiliary information in the individual surveys. The procedure is interesting from both coherence and accuracy points of view.

Furthermore, statistics may be related, although without clear connections in terms of, for example, units. Labour market statistics based on business surveys and on household surveys provide an example, see also Section 10.4.

10.3.4 Comparability over time

There is usually interest both in recent statistics and in long time series. Accuracy of changes is often at least as important as accuracy of levels.

Stability of definitions is important, but changes in structure should also be taken into account. For example, the use of chain-linked indices has increased, and an index with a fixed base is recommended to be rebased fairly frequently, at least every fifth year. It may be necessary to change variables to be in line with accounting practices if these change. New administrative rules may influence the BR in a way that carries over to the statistics. Comparability over time should be taken into account when choosing variables: current prices are often complemented with constant price or volume measures.

The methodology used has an influence on comparability, and there has to be a compromise between introducing for example improved estimation methods and keeping the old way with regard to time series. There is often a "jump" in a time series when a change is made. Hence, care is needed when introducing changes in methods. It may be wise to have a "double-run" period, that is to run the two methods in parallel to measure the effects and possibly link the two time series. As a minimum, explanations should be provided to the users.

When comparing short-term statistics, calendar and seasonal adjustments are important tools, with regard to corresponding periods in different years and adjacent periods. There are different methods of adjustment, building on different assumptions, like additive or multiplicative components. The appropriateness of a method is not necessarily the same in all countries. Still, for comparability reasons there should be some harmonisation of the adjustments of time series.

10.3.5 International comparability

As already indicated above in Section 10.3.1, there are many international harmonisation activities to improve comparability between countries.

Standard classifications is a typical example, with for example NACE Rev. 1 for classification of economic activities. There is a Regulation on Structural Business Statistics that includes definitions of variables. A Regulation on Short-Term Statistics has become law during 1998. There is a Regulation on statistical units – like enterprise, kind-of-activity unit and local unit – and also one on Business Registers. These regulations aim at increasing the comparability through making basic definitions equal – and also the applications similar by providing not only theory but also manuals with examples.

However, the subsidiarity approach means that each Member State may implement surveys in its own way, even when there is a regulation such as those mentioned. Similarly, regulations on for example statistical units may be interpreted and applied somewhat differently between MSs due to different traditions, prerequisites, etc. There are inherent cultural differences, like the number of working hours per full-time and part-time employee, the distribution of working hours over the year and over the week, taxation rules etc. The variable investments in fixed assets provides an example where the precise definition of the variable may vary between countries, at least before regulations have come into use. In such a case, it may be possible to make some kind of estimate as to the effect of a different national definition in comparison with the European concept. That is an attempt to overcome the lack in comparability, but to measure the difference is a difficult task.

Among examples of methods in the direction of comparability, standardisation of death rates in population demography is an old and illustrative one. Depoutot & Arondel (1997) discuss business statistics, and they advocate econometric models. Dalén (1998) presents sources of non-comparability in a general approach to the case of consumer price indices, and he presents empirical analyses of the effects of different conceptual and technical differences based on Swedish and Finnish data.

10.3.6 Some producer-based concluding comments

The discussions above and below show national and international actions to improve coherence including comparability, but also examples where deficiencies still remain. Many classification systems and regulations work in the direction of coherence between statistics from different surveys. Still, there are several classification systems. This means for instance that statistics on production of commodities and foreign trade statistics are difficult to use together when the former is based on PRODCOM and the latter on CN8. This influences for example the Producer Price Index (through the weights used for price indices for the domestic market, export etc) and the National Accounts.

The SBS and STS Regulations have much in common. There may still be deficiencies in coherence between annual and short-term statistics. One reason for differences is that these statistics partly build on different units, enterprise for annual statistics and kind-of-activity unit for short-term. Moreover, the latter uses kind-of-activity unit for example for manufacturing but, at least at present, enterprise for certain industries, for example services. The population is not clearly expressed for the STS, and the mixture of units seems to involve different practices, leading to further coherence and comparability deficiencies, with manufacturing kind-of-activity units within non-manufacturing enterprises and vice versa.

Another reason for differences between the two sets of statistics is different time schemes of production for the statistics for a given reference year. The annual statistics are collected after the year, while the short-term statistics are collected during the year. The population being surveyed changes during the year; births and deaths, mergers and break-ups etc. Such changes are better known when producing the annual than the short-term statistics.

Hence, even if the target populations are the same, the frames and the knowledge available may be different for the two surveys. That may imply differences – perhaps above all for the accuracy – that the producer should inform the users about. Alternatively, the producer may either revise the short-term statistics or refrain from using new population information for the annual statistics. This is an example of different practices in different Member States. See also Chapter 5.

The National Accounts build first on the short-term statistics. Later, when the annual statistic are available, both annual and short-term statistics are integrated into the accounts. The annual and the quarterly accounts are to be consistent, and so are the different accounts, like production and use. The National Accounts are to cover the whole economy. The integration may imply coherence deficiencies with both the annual and the short-term statistics, and, above all, inconsistent estimates.

A further example where coherence is interesting is between official short-term statistics and related statistics from other, possibly private, institutes; the latter may be qualitative, a barometer survey or business tendency survey.

As stated, it is important for the user to know if definitions are equal, or – if they are not – what the differences are. The differences should preferably be expressed in terms of effects on the statistics. The more accurate the statistics, the better in the joint use. Accurate statistics cannot, however, overcome different definitions. The user may find it convenient if estimation procedures are such that consistent estimates are obtained. The producer should consider these aspects when producing and presenting the statistics.

10.4 Some illustrations of coherence and co-ordination

As stated several times above, definitions are fundamental for coherence, including comparability. Accuracy is important, but in a different dimension. The more accurate the statistics, the stronger the inferences which can be made in the joint use. Random variation (for example due to sampling) is often easier to measure and take into account than systematic deviation (for example due to nonresponse) that is feared to be there, although difficult to quantify. If there are systematic deviations, it is easier to make comparisons if the deviations have a pattern that is stable.

In general, the closer the surveys, the less the problems with deficiencies in coherence. It is, however, neither possible nor desirable to have just one or a few surveys. There is a balance between “directed” surveys with few variables on the one hand and surveys with a broad scope and many variables on the other. The former way may allow comparatively small samples, but it may be convenient to include some variable – like the number of employees – in each survey. That means that the same variable value is reported many times. This increases the response burden. The system chosen should include willingness to respond and try to keep response burden low and spread out.

Co-ordination activities are important when several surveys are equal or at least similar. Germany is a notable example. Many surveys are performed on a sub-national level – in 16 ‘Bundesländer’ (regions) – and it is important to co-ordinate these surveys to obtain statistics not only for each ‘Bundesland’ but also on the federal level. There have to be compromises since optimal solutions are different depending on the level. For example, a good sample allocation for Germany may be quite different to that of individual ‘Bundesländer’. There is much to co-ordinate: variables, instructions, questionnaires, editing etc. Spain provides a similar example; 50 Provinces perform the initial data collection and editing.

There are related coherence problems in most countries. A survey may comprise all industries, or it may be more convenient to perform surveys for manufacturing and service industries separately. Annual and short-term surveys may be more or less co-ordinated as to variables. Often the annual survey is more detailed. A variable like turnover, or salaries and wages, may be included in both cases. There is an argument that for units in both samples it is unnecessary to ask for the sum of twelve values already collected, even if some of these are imputed. On the other hand, if monthly and annual data are collected, the annual survey will have problems with inconsistency if there has been some missing period, but if imputation is reasonable this inconsistency may be small. Moreover, a major aim of short-term surveys is to produce estimates quickly. If respondents do not have final results available for the month, they may be encouraged to provide estimates (their informed judgement being better than imputing for nonresponse). The source of the data for short-period and structural surveys may be different. The former may emanate from management or operational accounts. The latter are likely to be produced from the final audited accounts for the year and may include some adjustments which are made at the end of the year. On balance most countries see strong arguments for separate annual and short-period data collection.

Several countries now have one BR that is used as frame for all, or at least most, surveys. If all surveys use that also for updates, that will make the joint use of the statistics easier.

Several countries have introduced co-ordination of the sampling. There may be one or more aims: positive co-ordination to improve accuracy over time or between surveys and negative co-ordination (rotation is a possibility) mainly to spread the response burden.

There is a tension between annual statistics being as accurate as possible and being coherent with short-term statistics. Until recently in the UK coherence was the main driving force with annual panels selected to be consistent with short-term statistics. However, the emphasis has now switched to accuracy with the aim that the structural surveys should use the most up to date information available on units and classification. This change in policy means estimates closer to the target but larger revisions when benchmarking short-term statistics on annual statistics. Such a practice has a longer history of use in Sweden.

There may be co-ordination between surveys to ensure that the final statistics agree. There are different techniques depending on information and “closeness” of surveys. In Sweden, for example, the short-term index of production for the manufacturing industries is benchmarked on the annual index in spite of there being some differences in definitions; the short-term index being regarded as an indicator of the annual index, see the Swedish Model Quality Reports for descriptions and figures[6]. In the UK the short-term production index is not currently benchmarked to the annual surveys (but benchmarking is undertaken elsewhere). However, the UK strategy for the longer term is to move to chain linking supported by annual input-output tables. A consequence is that the value added from the annual surveys will replace estimates of gross output used in the short term. In this approximation a necessary assumption is that the ratio of gross to net output is constant over time. That hypothesis is

[6] This benchmarking has been debated in Statistics Sweden, and in late 1998 it was decided to discontinue.

likely to become stretched, particularly at lower levels of the SIC, the further one moves from the base year.

In many cases, different definitions may be found impossible to overcome and important to use for each of the single surveys. There may for example be different sets of employment statistics and of statistics on salaries and wages, coming from labour force surveys and business surveys, and from business surveys and administrative data tied to employers' declarations. A UK experience is that one way of helping users to understand the differences between the labour force and the employment surveys is to emphasise the differences between *people* and *jobs*, see Pease (1997). Making this distinction clear has helped to prevent users from focusing on the differences between the estimates which, when sampling errors are taken into account, are relatively small. Similarly, considerable resources have been used in the Netherlands on statistical integration for the labour market, with statistics based on establishment surveys, household surveys, and central registers, Leunis & Altena (1996).

The co-ordination may be on the macro level, as just mentioned, and/or on the micro level. There may for example be an exchange of figures for individual units between surveys perhaps to ensure that an enterprise that is complex and/or re-organising is fully included or as a part of the editing system. There are such practices in Germany and Denmark. Similarly, staff in the UK generally work on more than one inquiry. In the production sector the same data collector will work on Stocks, Capital Expenditure, Monthly Turnover and the Annual Structural Surveys. Thus comparisons of data at contributor level can easily be made and actions taken to reconcile differences.

Member States often make changes to their inquiry systems to improve the methodology and achieve greater consistency with other surveys, classifications or European regulations. Although these developments may increase coherence between surveys and countries, they introduce discontinuities when the changes are made – distorting the comparability over time. Specific examples of changes that influence definitions and/or accuracy include:

(a) changes of administrative rules or data, for example data used for updates, regional boundary changes
(b) construction of a new register or frame
(c) new sampling design
(d) changes in estimation methodology
(e) new outlier treatment
(f) move to NACE Rev. 1
(g) move to ESA (European System of Accounts)

In order to calculate a link between the two time series, it is necessary to have statistics on both the old and the new basis. There is analytical work and often extra data collection. Nonetheless, the work is vital since the link factors are often large even for changes which may seem to be slight. For example in the UK changes in estimation methods have at times altered industry totals at class level by over ten per cent. The links may be calculated for a month, a quarter, or a number of periods. Where links are large and could vary from period to period it may be best to look at some average link over a period of time to ensure stability. Any cases where the factor is surprisingly large or small should be followed up. Links can be applied to either the old or the new series.

Part 4: Conclusions and References

11 Concluding remarks

Paul Smith, Office for National Statistics

11.1 Methodology for quality assessment

This volume contains a lot of information on the theory and methods behind the assessment of quality in business surveys, covering a huge range of techniques. In many survey situations it will be practical only to use a small number of these to assess the quality of the survey results, because of the limitations of time, money and available information. A natural choice is to aim for a balance between those methods which are easy to apply and evaluation of the quality components which are the most important ones.

Some accuracy measures have a long tradition, for example the sampling error when the sample design is probability-based. Often these measures are those most amenable to theoretical treatment. Software for assessing the sampling errors is reviewed in Volume II, and the properties of sampling errors are also investigated there.

Non-sampling errors and non-probability sampling schemes are accessible to investigation by three main general methods:

- indicators;
- follow-up studies; and
- sensitivity analysis.

Indicators are statistics, normally available as by-products of the survey processing, which are thought to be (strongly) correlated with the quality of the estimates, but which do not *directly* measure that quality. They are the easiest statistics on quality to calculate, and they predominate in the model quality reports (volume III), although the precise details differ according to what needs to be estimated. Both follow-up studies and sensitivity analysis are limited by the data which are available (or obtainable); follow-up studies are typically high-cost (for NSIs and contributors) but aim to get closer to the true value than the original survey did, usually for a subset of the original observations. Sensitivity analyses rely on the data already available (both survey and auxiliary data) to suggest plausible models, and indicate how the estimates change with different models (or different assumptions). In a small number of cases NSIs obtain "follow-up" data as part of the survey process, and need only insert some extra storage or undertake some additional work to use it – in particular processing error and coding error (where all the original responses are available (if they are stored) and can be re-evaluated) and nonresponse error (where the change of response with time gives some idea of the characteristics of nonrespondents). In general however, follow-up studies are detailed and very expensive, and are undertaken rarely and on a small scale.

Sensitivity analysis is cheaper as it uses only the data already collected and requires only the reprocessing of this data under different scenarios. It gives an indication of how the estimate is affected by certain models and assumptions, but does not say how close these estimates are to the true value, although there is an implicit assumption that if all the scenarios investigated have similar outcomes, then these outcomes will be close to the true value.

Deducing which components of total survey error (see section 1.2.1) have the biggest contribution is much more problematic, since in different surveys the answer may well be different, and there is only a small number of studies which investigates several errors in a single survey in a comparative way. It is perceived wisdom that "non-sampling errors may outweigh sampling errors substantially", but there is little evidence of the relative importance of errors in practice. Much of the methodology behind survey estimation involves on the one hand removing bias as much as possible and accepting an increase in variance (for example in compensating for nonresponse, chapter 8), and on the other hand introducing bias in a structured way to reduce the variability of survey estimates (for example through poststratification or outlier adjustment), so it is measurement of these biases and variances which will lead to the total survey error.

11.2 Recommendations for quality assessment

Clearly it is inappropriate to undertake an in-depth study of all the biases and variance components of a survey on every occasion that it is run. However, it is also clear that this sort of study is the only way in which a complete evaluation of the survey quality can be made. This leads us to suggest a three-pronged approach to evaluating quality:

(a) Indicators should be included as part of survey processing systems, and should be produced each time the survey is run. They not only indicate the quality, but also show where survey processes are failing. These should include for instance weighted and unweighted response rates, rate of identification of misclassifications and dead units, and data edit failure rates.

(b) Quality measures should be produced periodically (at least annually) where they are clearly defined. These should include sampling errors.

(c) There should be a rolling programme of evaluation of the overall quality of the survey, covering some topics each year. This would involve the use of follow-up interviews and other detailed studies, in order to estimate the true total survey error. The exact list of components to be included would need to be decided; ideally all components would be measured. Some of the burden of measurement could be moved away from the survey by, for example, undertaking an evaluation of the frame quality, as the frame is used for many surveys.

In addition to these three, a useful qualitative measure of survey quality is to have the methods fully documented, and to have the quality assessment practices written down, much as in the Model Quality Reports. The act of producing these reports will force the methods of the survey to be considered critically, and this will influence the quality.

The Model Quality Reports (volume III) include both simple indicators and more ambitious measures like sensitivity analyses, but not in-depth studies. The Implementation Reports and the Guidelines on implementation (volume IV) include discussions of balancing issues.

12 References

ABRAHAM, B. & LEDOLTER, J. (1983) *Statistical methods for forecasting*. New York: Wiley.

ARCHER, D. (1995) Maintenance of business registers. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 85-100. New York: Wiley.

BETHLEHEM, J.G. (1988) Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics,* **4**, 251 - 260.

BERGER, Y.G. (1998) Rate of convergence to asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **72**, 149-168.

BIEMER, P.P. & FECSO, R.S. (1995) Evaluating and controlling measurement error in business surveys. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 257 – 281. New York: Wiley.

BLOM, E. & FRIBERG, R. (1995) The use of scanning at Statistics Sweden. *Proceedings of the International Conference on Survey Measurement and Process Quality, Contributed papers,* pp 52-63. Virginia: American Statistical Association.

BOX, G.E.P., JENKINS, G.M. & REINSEL, G.C. (1994) *Time series analysis, forecasting, and control*, third edition. Englewood Cliffs, NJ: Prentice-Hall.

BUSHNELL, D. (1996) Computer-assisted occupation coding. In *Proceedings of the Second ASC International Conference* (eds. R. Banks, J. Fairgrieve, L. Gerrard, T. Orchard, C. Payne, & A. Westlake), pp 165-173. Chesham: Association for Survey Computing.

CANTY, A.J. & DAVISON, A.C. (1997) *Variance estimation for the Labour Force Survey*. Report prepared under contract to the University of Essex, on behalf of the UK Office for National Statistics.

CENTRAL STATISTICAL OFFICE (1992) *Standard industrial classification of economic activities 1992*. Newport: CSO.

CHAMBERS, R.L. (1986) Outlier robust finite population estimation. *Journal of the American Statistical Association*, **81**, 1063-1069.

CHAMBERS, R.L. (1997) Weighting and calibration in sample survey estimation. In *Proceedings of the Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth* (eds. C. Malaguerra, S. Morgenthaler & E. Ronchetti), Monte Verità, Switzerland, Basel: Birkhäuser Verlag.

CHAMBERS, R.L. (1997) *Small-area estimation: a survey sampler's perspective*. Technical report, Department of Social Statistics, University of Southampton, UK (presented at a meeting organized by the Small Area Health Statistics Unit of Imperial College, UK, May 1997).

Chambers, R.L. & Dorfman, A.H. (1994) Robust sample survey inference via bootstrapping and bias correction: the case of the ratio estimator. *Invited Paper, Joint Statistical Meetings of the ASA, IMS and the Canadian Statistical Society, Toronto, August 13-18, 1995.*

Chambers, R.L., Dorfman, A.H. & Wehrly, T.E. (1993) Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, **88**, 268-277.

Chambers, R.L. & Kokic, P.N. (1993) Outlier robust sample survey inference. *Invited Paper, Proceedings of the 49th Session of the International Statistical Institute, Firenze, August 25-September 2, 1993.*

Chatfield, C. (1996) *The analysis of time series: an introduction*. London: Chapman & Hall.

Cholette, P.A. & Dagum, E.B. (1994) Benchmarking time series with autocorrelated survey errors. *International Statistical Review*, **62**, 365-377.

Cochran, W .G. (1977) *Sampling techniques*, third edition. New York: Wiley

Cook, R.D. & Weisberg, S. (1982) *Residuals and influence in regression*. London: Chapman & Hall.

Copas, J.B. & Li, H.G. (1997) Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B,* **59***,* 55-96

Cowling, A., Chambers, R., Lindsay, R. & Parameswaran, B. (1996) Applications of spatial smoothing to survey data. *Survey Methodology*, **22**, 175-183.

Cox, D.R. & Hinkley, D.V. (1974) *Theoretical statistics*. London: Chapman & Hall.

Dagum, E.B., Cholette, P.A. & Chen, Z.-G. (1998) A unified view of signal extraction, benchmarking, interpolation and extrapolation in time series. *International Statistical Review*, **66**, 245-269.

Dalén, J. (1998) Studies on the comparability of consumer price indices. *International Statistical Review*, **66**, 83-113.

Dawid, A.P. (1979) Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1-31.

Dawid, A.P. & Lauritzen, S.L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272-1317.

Deming, W.E. (1956) On simplification of sample design through replication with equal probabilities and without stages. *Journal of the American Statistical Association*, **51**, 24-53.

Denton, F.T. (1971) Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, **66**, 99-102.

DEPOUTOT, R. & ARONDEL, PH. (1997) International comparability and quality of statistics. *Presented to CAED97, International Conference on Comparability Analysis of Enterprise (micro)Data, 15-17 December 1997, Bergamo, Italy.*

DEVILLE, J.-C. (1991) A theory of quota surveys. *Survey Methodology*, **17**, 163-181.

DEVILLE, J.-C. & SÄRNDAL, C.-E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

DEVILLE, J.-C. & SÄRNDAL, C.-E. (1994) Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, **10**, 381-394.

DIPPO, C.S., CHUN, Y.I. & SANDER, J. (1995) Designing the data collection process. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 283-301. New York: Wiley.

DRAPER, D. (1995a) Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 45-97.

DRAPER, D. (1995b) Inference and hierarchical modelling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics*, **20**, 115-147, 233-239.

DRAPER, D. & CHEAL, R. (1998) *Causal inference via Markov chain Monte Carlo*. Technical report, Statistics Group, Department of Mathematical Sciences, University of Bath, UK.

DRAPER, D., GAVER, D., GOEL, P., GREENHOUSE, J., HEDGES, L., MORRIS, C., TUCKER, J. & WATERNAUX, C. (1993a) *Combining information: statistical issues and opportunities for research*. Contemporary Statistics Series, No. 1. Alexandria VA: American Statistical Association.

DRAPER, D., HODGES, J., MALLOWS, C. & PREGIBON, D. (1993b) Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, **156**, 9-37.

DURBIN, J. & QUENNEVILLE, B. (1997) Benchmarking by state space models. *International Statistical Review,* **65**, 23-48.

EFRON, B. & TIBSHIRANI, R.J. (1993) *An introduction to the bootstrap*. London: Chapman & Hall.

ELDER, S. & MCALEESE, I. (1996) Application of document scanning, automated data recognition and image retrieval to paper self-completion questionnaires. In *Survey and Statistical Computing* 1996. [Eds note: full ref please]

ELVERS, E. (1993) A new Swedish business register covering a calendar year and examples of its use for estimation. *Proceedings of the International Conference on Establishment Surveys*. American Statistical Association, pp. 916-919.

EUROSTAT (1996:04) *Proposal for a quality report on structural business indicators* Eurostat/D3/Quality/96/04. Luxembourg: Eurostat.

EUROSTAT (1997:04) *Proposal for a quality report on short-term indicators*. Eurostat/A4/Quality/97/04. Luxembourg: Eurostat.

EUROSTAT (1997:06) *Variance estimation of static statistics. Part 1: Overview}*. Eurostat/A4/Quality/97/06. Luxembourg: Eurostat.

EUROSTAT (1997:07) *Variance estimation for dynamic statistics. A simulation study* (draft). Eurostat/A4/Quality/97/07. Luxembourg: Eurostat.

EUROSTAT (1998:07) *Methodological aspects of producer prices on the export market*. Eurostat/4E/Energy and Industry/98/07. Luxembourg: Eurostat.

EUROSTAT (1998a) *Quality of Business registers.* Eurostat/D1/BRSU/98-12 (Working group meeting 29-30 June 1998).

EUROSTAT (1998b) *Seasonal adjustment methods: a comparison*. Statistical Document, Theme 4E, 1998. Luxembourg: Eurostat.

FAY, R.E. (1996) Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association,* **91**, 490-498.

FELLEGI, I.P. & HOLT, D. (1976) Systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, **71**, 17-35

FINDLEY, D.F., MONSELL, B.C., BELL, W.R., OTTO, M.C. & CHEN B.-C. (1998) New capabilities and methods of the X12-ARIMA seasonal adjustment program (with discussion). *Journal of Business and Economic Statistics*, **16**, 127-177.

FISK, P.R. (1977) Some approximations to an “ideal” index number. *Journal of the Royal Statistical Society, Series A*, **140**, 217–231.

FORSTER, J.J. & SMITH, P.W.F. (1998) Model-based inference for categorical survey data subject to non-ignorable nonresponse (with discussion). *Journal of the Royal Statistical Society, Series B*, **60**, 57-70, 89-102.

FREEDMAN, D., PISANI, R. & PURVES, R. (1998) *Statistics*, third edition. New York: Norton.

FRIBERG (1992) Surveys on environmental investments and costs in Swedish industry. *Statistical Journal of the UN Economic Commission for Europe,* **9**, 101-110.

GHOSH, M. & RAO, J.N.K. (1994) Small-area estimation: an appraisal (with discussion). *Statistical Science*, **9**, 55-93.

GLYNN, R.J., LAIRD, N.M. & RUBIN, D.B. (1986) Selection modeling versus mixture modeling with non-ignorable nonresponse. In *Drawing inferences from self-selected samples* (ed. H. Wainer), pp. 115-152. New York: Springer.

GOLDSTEIN, H., RASBASH, J., PLEWIS, I., DRAPER, D., BROWNE, B., YANG, M. & WOODHOUSE, G. (1997) *A User's Guide to ML*n *for Windows*™ (ML*wi*N), *Version 1.0b, November 1997*. London: Institute of Education.

GOMEZ, V. & MARAVALL, A. (1994a) *Program SEATS (Signal Extraction in ARIMA Time Series): Instructions for the User*. Working Paper ECO 94/28, European University Institute, Florence.

GOMEZ, V. & MARAVALL, A. (1994b) *Program TRAMO (Time series Regression with ARIMA noise, Missing Observations, and Outliers): Instructions for the User*. Working Paper ECO 94/31, European University Institute, Florence.

GRANQUIST, L. (1984) On the role of editing. *Statistical Review,* **2**, 105-118.

GRANQUIST, L. (1995) Improving the traditional editing process. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 385-401. New York: Wiley.

GRANQUIST, L. & KOVAR, J.G. (1997) Editing of survey data: how much is enough? In *Survey Measurement and Process Quality* (eds L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin), pp. 415-435. New York: Wiley.

GRIFFITHS, G. & LINACRE, S. (1995) Quality assurance for business surveys. . In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 673-690. New York: Wiley.

GROVES, R.M. (1989) *Survey errors and survey costs*. New York: Wiley.

HAAN, J. DE, OPPERDOES, E. & SCHUT, C. (1997) Item sampling in the consumer price index: a case study using scanner data. Paper submitted to the *Joint ECE/ILO Meeting on Consumer Price Indices* (Geneva, 24-27 November 1997).

HÀJEK, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, **35**, 1491-1523.

HAMMERSLEY, J.M. & HANDSCOMB, D.C. (1979) *Monte Carlo methods*. London: Chapman & Hall.

HARVEY, A.C. (1989) *Forecasting, structural time series models, and the Kalman filter*. Cambridge: Cambridge University Press.

HECKMAN, J.J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153-161.

HIDIROGLOU, M.A. & BERTHELOT, J.M. (1986) Statistical editing and imputation for periodic business surveys. *Survey Methodology*, **1**, 73-83

HIDIROGLOU, M.A., SÄRNDAL, C.-E. & BINDER, D.A. (1995) Weighting and estimation in business surveys. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 477-502. New York: Wiley.

HILLMER, S.C. & TRABELSI, A. (1987) Benchmarking of economic time series. *Journal of the American Statistical Association*, **82**, 1064-1071.

HOLT, D. & SMITH, T.M.F. (1979) Post-stratification. *Journal of the Royal Statistical Society, Series A*, **142**, 33-46.

HORVITZ, D.G. & THOMPSON, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

HUBER, P.J. (1981) *Robust statistics*. New York: Wiley.

JAGERS, P. (1986) Post-stratification against bias in sampling. *International Statistical Review*, **54**, 159-167.

JAZAIRI, N.T. (1982) Index numbers. Entry in *Encyclopedia of Statistical Sciences*, **4** (eds. N.L. Johnson & S. Kotz). New York: Wiley.

JONES, R.H. (1980) Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, **22**, 389-395.

KALTON, G. & STOWELL, R. (1979) A study of coder variability. *Journal of the Royal Statistical Society, Series C*, **28**, 276-289.

KASPRZYK, D. & KALTON, G. (1998) Measuring and reporting the quality of survey data. *Symposium '97, New Directions in Surveys and Censuses: proceedings* pp. 179-184. Ottawa: Statistics Canada.

KENNY, P.B. & DURBIN, J. (1982) Local trend estimation and seasonal adjustment of economic and social time series (with discussion). *Journal of the Royal Statistical Society, Series A*, **145**, 1-45.

KOHN, R. & ANSLEY, C.F. (1986) Estimation, prediction, and interpolation for ARIMA models with missing observations. *Journal of the American Statistical Association*, **81**, 751-761.

KOKIC, P.N. (1998) Estimating the sampling variance of the UK Index of Production. *Journal of Official Statistics*, **14**, 163-179.

KOKIC, P.N. & SMITH, P.A. (1999a) Winsorisation of outliers in business surveys. Submitted to *Journal of the Royal Statistical Society, Series D*.

KOKIC, P.N. & SMITH, P.A. (1999b) Outlier-robust estimation in sample surveys using two-sided winsorisation. Submitted to *Journal of the American Statistical Association*.

KOVAR, J.G. &WHITRIDGE, P.J. (1995) Imputation of business survey data. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 403-423. New York: Wiley.

LEE, H. (1995) Outliers in business surveys. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 503-526. New York: Wiley.

LESSLER, J.T. & KALSBEEK, W.D. (1992) *Non-sampling errors in surveys*. New York: Wiley.

LEUNIS, W.P. & ALTENA, J.W. (1996) Labour accounts in the Netherlands, 1987-1993. How to cope with fragmented macro data in official statistics. *International Statistical Review,* **64**, 1-22.

LITTLE, R.J.A. (1993) Post-stratification: a modeller's perspective. *Journal of the American Statistical Association*, **88**, 1001-1012.

LUNDSTRÖM, S. (1997) *Calibration as a standard method for treatment of nonresponse*. Doctoral dissertation, Department of Statistics, University of Stockholm.

LYBERG, L. & KASPRZYK, D. (1997) Some aspects of post-survey processing. In *Survey Measurement and Process Quality* (eds L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin), pp. 353-370. New York: Wiley.

MAHALONOBIS, P.C. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, **109**, 325-370.

MARAVALL, A. (1998) Comment on "New capabilities and methods of the X12-ARIMA seasonal adjustment program," by Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.-C.. *Journal of Business and Economic Statistics*, **16**, 155-160.

MOSTELLER, F. & TUKEY, J.W. (1977) *Data analysis and regression*. Reading, MA: Addison-Wesley.

NASCIMENTO SILVA, P.L.D. & SKINNER, C.J. (1997) Variable selection for regression estimation in finite populations. *Survey Methodology,* **23**, 23-32.

NEYMAN, J. (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558-606.

NORDBERG, L. (1998) *On variance estimation for measures of change when samples are co-ordinated by a permanent random number technique*. R&D Report 1998:6, Statistics Sweden.

OHLSSON, E. (1995) Coordination of samples using permanent random numbers. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 153-169. New York: Wiley.

PEASE, P. (1997) Comparison of sources of employment data. *Labour Market Trends*, December 1997. London: Office for National Statistics.

PIERZCHALA, M. (1990) A review of the state of the art in automated data editing and imputation. *Journal of Official Statistics*, **6**, 355-377.

PURCELL, N.I. & KISH, L. (1980) Post-censal estimates for local areas (or domains). *International Statistical Review*, **48**, 3-18.

RAO, J.N.K. (1996) On variance estimation with imputed survey data. *Journal of the American Statistical Association ,* **91**, 499-520.

RENSSEN, R.H. & NIEUWENBROEK, N.C. (1997) Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association,* **92**, 368-374.

ROYALL, R.M. (1982) Finite populations, Sampling from. Entry in the *Encyclopedia of Statistical Sciences* (eds. N.L. Johnson & S. Kotz). New York: Wiley.

ROYALL, R.M. (1986) The prediction approach to robust variance estimation in two-stage cluster sampling. *Journal of the American Statistical Association*, **81**, 119-123.

ROYALL, R.M. & CUMBERLAND, W.G. (1981) An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, **76**, 66-88.

ROYALL, R.M. & HERSON, J. (1973) Robust estimation in finite populations I. *Journal of the American Statistical Association*, **68**, 880-889.

RUBIN, D.B. (1986) Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, **12**, 37-47.

RUBIN, D.B. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association,* **91***,* 473-489.

SÄRNDAL, C.-E. (1992) Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology,* **18**, 241-252.

SÄRNDAL, C.-E. & SWENSSON, B. (1987) A general review of estimation for two phases of selection with applications to two-phase sampling and non-response. *International Statistical Review*, **55**, 279-294.

SÄRNDAL C.-E., SWENSSON B. & WRETMAN, J. (1992) *Model-assisted survey sampling*. New York: Springer-Verlag.

SEN, A.R. (1953) On the estimation of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119-127.

SHAO, J. & TU, D. (1995) *The jackknife and bootstrap*. New York: Springer-Verlag.

SKENE, A.M., SHAW, J.E.H. & LEE, T.D. (1986) Bayesian modeling and sensitivity analysis. *The Statistician*, **35**, 281-288.

SMITH T.M.F. (1983) On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society, Series A*, **146**, 394-403.

SMITH T.M.F. (1991) Post-stratification. *The Statistician*, **40**, 315-323.

SMITH T.M.F. (1993) Populations and selection - limitations of statistics. *Journal of the Royal Statistical Society, Series A*, **156**, 145-166.

SOS (1998) *New started enterprises in Sweden 1996 and 1997*. Statistical Report Nv 12 SM 9801 in the series *Official Statistics of Sweden*. Örebro, Sweden: Statistics Sweden.

STATISTICS FINLAND (1996) Progress Report. Contribution by T. Viitaharju and A. Heinonen to the *10th International Roundtable on Business Survey Frames*.

STATISTICS SWEDEN (1995) Demography of enterprises and establishments in Sweden. An employment approach to measuring the dynamics among units. Contribution by B. Tegsjö to the *9th International Roundtable on Business Survey Frames*. SCB, Örebro, pp. 255-262.

STRUIJS, P. & WILLEBOORDSE, A. (1995) Changes in populations of statistical units. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 65-84. New York: Wiley.

SUGDEN R.A. (1993). Partial exchangeability and survey sampling inference. *Biometrika*, **80**, 451-455.

THEIL, H. (1960) Best linear index numbers of prices and quantities. *Econometrica*, **28**, 464-480.

VEZINA, S. (1996) Statistics Canada's experiences with automated data entry. In *Proceedings of Statistics Canada's Symposium 96*, Ottawa.

WEISBERG, S. (1985) *Applied regression analysis*, second edition. New York: Wiley.

WOLTER, K.M. (1985) *Introduction to variance estimation*. New York: Springer-Verlag.

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

YATES, F. & GRUNDY, P.M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*, **15**, 253-261.

13 Index