Southampton Statistical Sciences Research Institute
University of Southampton

Which Sample Survey Strategy? A Review of Three Different Approaches

R. L. Chambers

Abstract

Sample survey theory is concerned with methods of sampling from a finite population of N units and then making inferences about finite population quantities on the basis of the sample data. A method of sampling coupled with a method of estimation given the sample data is often referred to as a sampling strategy, and typically corresponds to a set of rules which tell one how to obtain a sample of units from the finite population and then how to manipulate the resulting sample data to estimate the value of a quantity defined for the entire population.

S^{3}RI Methodology Working Paper M03/20

Which Sample Survey Strategy? A Review of Three Different Approaches

R. L. Chambers

Southampton Statistical Sciences Research Institute

University of Southampton

Revised, September 2003

1. Introduction

Sample survey theory is concerned with methods of sampling from a finite population of N units and then making inferences about finite population quantities on the basis of the sample data. A method of sampling coupled with a method of estimation given the sample data is often referred to as a sampling strategy, and typically corresponds to a set of rules which tell one how to obtain a sample of units from the finite population and then how to manipulate the resulting sample data to estimate the value of a quantity defined for the entire population.

In this paper we review the essential characteristics of the three different approaches to specifying a sampling strategy; the design based approach, the model assisted approach and the model based approach. All three approaches are in use in major statistical agencies. Furthermore, the advantages and disadvantages of all three have been hotly debated in the sampling theory literature in recent years. See the sequence of papers Smith (1976), Smith (1984) and Smith (1994) for a clear and entertaining description of how this debate has progressed. However, with the exception of Brewer and Särndal (1983), there seems to have been little attempt to view all three approaches from a unified statistical perspective.

We describe such a unified framework for survey design and estimation below. After embedding the above three approaches in this common framework, we then contrast them in terms of their concepts of efficiency as well as their robustness to assumptions about the characteristics of the finite population. Our conclusion is that although no one approach

delivers both efficiency and robustness, the model-based approach seems to achieve the best compromise between these typically conflicting objectives.

The paper is organised as follows. In the following section, some basic concepts of the statistical (as opposed to the practical) theory of sample design and estimation are introduced, along with the common distributional framework that underlies the development in the paper. In sections 3 through 5 this framework is used to characterise the statistical basis for choosing an efficient sampling strategy under the above three approaches. The paper concludes in section 6 with a discussion of how the different robustness concepts that apply within each approach relate to one another within this framework, together with comments on the strengths and weaknesses of each.

2. Basic Concepts

A fundamental sample survey concept is that of a population frame. This is the list of the N units making up the finite population. Availability of a frame is a basic requirement underlying most (though not all) of sample survey theory, since sampling methods are usually expressed in terms of rules for deciding which elements of a list constitute the sample. Note that this frame requirement does not exclude multistage surveys from consideration, since these require access to an initial frame for selection of first stage units and then subsequent "sub-frames" for selection of second and later stage units.

In what follows it will be assumed that the frame always contains a unique identifier or label for each unit of the population. In many cases the frame also contains the values of one or more auxiliary variables associated with each unit of the finite population. We use the index I (and sometimes J and K as well) to denote the population labels. Without loss of generality we can assume these labels take the values 1, 2, .., N, so it makes sense to refer to the I^{th} population unit.

For simplicity we also assume in this paper that there is only one auxiliary variable, X and one survey variable Y. We put X_I equal to the value of X for the I^{th} population unit, with Y_I defined similarly. The extension to multiple Y-variables and multiple X-variables does not involve introduction of new concepts, but does make the notation much more complex.

A traditional objective of most surveys is estimation of the finite population total,

$$T = \sum_{I=1}^{N} Y_I .$$

In order to estimate T, the survey sampler (who is assumed to have access to the population frame) "selects" a sample of units from the population by identifying their labels on the frame, and then measures their corresponding values of Y. These sample values of Y are then combined with framework information to generate the required estimate of T.

A convenient way of characterising the above sample selection process is to assume that, for each unit I on the frame, the survey sampler generates a "new" variable S_I which takes a value equal to the number of times that particular population unit's Y value is observed. The set of labels corresponding to population units that are sampled in this way is $s = \{I: S_I > 0\}$. The labels of the remaining non-sampled population units then define the set $r = \{I: S_I = 0\}$.

The distribution of these S_I values effectively defines what is generally referred to as the design of the sample survey. In principle at least, the survey sampler has complete control over (and hence complete knowledge of) this distribution. Let $\mathbf{S}$ denote the vector valued random variable corresponding to the N population values of S_I and let $\mathbf{X}$ denote the corresponding N-vector of population values of the X_I.

Throughout, we assume that the distribution of $\mathbf{S}$ only depends on the known population values in $\mathbf{X}$. That is, given $\mathbf{X}$, the distribution of $\mathbf{S}$ is completely specified. This assumption is often referred to as "uninformative sampling" in the literature (e.g. Pfeffermann 1993). An immediate consequence is that, given $\mathbf{X}$, the distributions of $\mathbf{S}$ and $\mathbf{Y}$, the population N-vector of Y values, are independent. Note that this assumption is not appropriate in cases where the sampler has limited or no control over the sampling process, since in those cases there is the possibility of selection bias, and consequent dependence between the distributions of $\mathbf{Y}$ and $\mathbf{S}$, even after conditioning on $\mathbf{X}$. See Smith (1983) and Sugden and Smith (1984).

This characterisation of the outcome of sample selection as a random vector $\mathbf{S}$ includes probability sampling as a special case. However, since there is no requirement that the distribution of $\mathbf{S}$ be non-degenerate, it also includes many so-called "non-random" (strictly speaking non-probability sampling based) selection methods used by survey practitioners.

Turning now to estimation of T, we see that any estimator of this quantity can only be based on the available data, i.e. the observed sample values of Y together with relevant frame information (the population values of X and S). Consequently, we consider a general linear estimator of T of the form

$$\hat{T} = \sum_{s} W_I(\mathbf{S},\mathbf{X}) S_I Y_I$$

where $W_I(\mathbf{S},\mathbf{X})$ is a "sample weight" associated with population unit I when this unit appears in sample. Observe that this weight generally depends on **S** and **X**. Also, we can equivalently write

$$\hat{T} = \sum_{I=1}^{N} W_I(\mathbf{S},\mathbf{X}) S_I Y_I$$

since $S_I = 0$ for non-sampled population units.

In general, we can think of the values of the $W_I(\mathbf{S},\mathbf{X})$ as characterising the estimation process in the same way that the values S_I characterised the sampling process. That is, the survey sampler can be considered as in principle defining a value $W_I(\mathbf{S},\mathbf{X})$ for each population unit (whether sampled or not), and consequently a "new" variable over the population. The value of this "weight variable" W may depend on the population values of the auxiliary variable, X, and those of the selection variable, S, but not those of the survey variable Y. That is, if we put **W** equal to the population vector defined by the $W_I(\mathbf{S},\mathbf{X})$, then **W** is a random vector whose distribution is completely determined by that of **S** and **X**. To keep our notation simple, we drop explicit referencing to **S** and **X** when writing down individual weights from now on, writing $W_I = W_I(\mathbf{S},\mathbf{X})$ in what follows. In all cases, however, the reader should keep in mind the dependence of these weights on the realised values of **S** and **X**.

A sampling strategy corresponds to the pair (**S**, **W**). Deciding on a sampling strategy therefore consists of (i) given **X**, choosing an appropriate distribution for **S**, and (ii) given **X** and the distribution generated under (i), choosing an appropriate specification for **W**. In the following sections we compare and contrast three different approaches to carrying out (i) and (ii) above. In particular, we focus on choice of an optimal strategy in each case.

Finally, we observe that in all cases, the inferential framework assumed is the one defined by the joint distribution of **S**, **X** and **Y**. That is, the sample space for inference is the one

corresponding to all possible realisations of these three vectors. Consequently, all relevant probabilities are defined with respect to this joint distribution, as are expectations and variances (both conditional and unconditional).

3. Optimality under the Design Based Approach

This approach has its origin in Neyman's key paper (Neyman, 1934). It also represents the basic underlying philosophy in most traditional sampling theory texts, e.g. Cochran (1977), Kish (1965). A key concept under this approach is that of design unbiasedness. That is, for any choice of sampling process S, the weighting process W must be such that the frequency weighted average value of $\hat{T}$ over all possible samples generated under S is the actual value of T. In other words, this approach restricts consideration to those weights **W** which ensure that, irrespective of the particular sample selection method (i.e. **S**) employed,

$$E\left(\hat{T} - T \middle| \mathbf{X}, \mathbf{Y}\right) = 0. \tag{1}$$

for all values of **Y** and **X**.

For (1) to hold for arbitrary **Y** and **X** we must have

$$E(W_I S_I \middle| \mathbf{X}, \mathbf{Y}) = 1$$

or, since the distributions of both **S** and **W** are completely determined by that of **X**,

$$E(W_I S_I \middle| \mathbf{X}) = 1.$$

A sufficient condition for this to be satisfied is clearly where

$$W_I^{-1} = E(S_I \middle| \mathbf{X}) \tag{2}$$

in which case

$$\hat{T} = \sum_s \frac{S_I Y_I}{E(S_I \middle| \mathbf{X})}. \tag{3}$$

The design based approach requires that all inferential probabilities be conditioned on both **Y** and **X**. Consequently, under this approach the efficiency of $\hat{T}$ is measured by

$$\mathrm{var}(\hat{T} - T \middle| \mathbf{X}, \mathbf{Y}) = \mathrm{var}(\hat{T} \middle| \mathbf{X}, \mathbf{Y}) = \mathrm{var}\left(\sum_{I=1}^{N} W_I S_I Y_I \middle| \mathbf{X}, \mathbf{Y}\right) = \sum_{I=1}^{N}\sum_{J=1}^{N} \mathrm{cov}(W_I S_I, W_J S_J \middle| \mathbf{X}) Y_I Y_J$$

which, in the case of the weighting method (2) above, becomes

$$\mathrm{var}(\hat{T} - T \middle| \mathbf{X}, \mathbf{Y}) = \sum_{I=1}^{N}\sum_{J=1}^{N} \frac{\mathrm{cov}(S_I, S_J \middle| \mathbf{X})}{E(S_I \middle| \mathbf{X}) E(S_J \middle| \mathbf{X})} Y_I Y_J. \tag{4}$$

In some circumstances (e.g. where a particularly complex sampling method has been employed) it may be impossible to evaluate (2) exactly. In such cases condition (1) is at least approximately true whenever

$$W_I^{-1} = \hat{E}(S_I \,|\, \mathbf{X}) \tag{5}$$

where $\hat{E}$ denotes an estimate of the conditional regression function, based on the frame values of *X*. In this case the bias of the resulting estimator $\hat{T}$ under the design based approach is

$$E\left(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}\right) = \sum_{I=1}^{N} \left(\frac{E(S_I \,|\, \mathbf{X})}{\hat{E}(S_I \,|\, \mathbf{X})} - 1 \right) Y_I \tag{6}$$

with variance

$$\text{var}(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) = \sum_{I=1}^{N} \sum_{J=1}^{N} \frac{\text{cov}(S_I, S_J \,|\, \mathbf{X})}{\hat{E}(S_I \,|\, \mathbf{X}) \hat{E}(S_J \,|\, \mathbf{X})} Y_I Y_J \,. \tag{7}$$

The measure of efficiency of $\hat{T}$ in this case is therefore its design mean squared error

$$\begin{aligned} MSE(\hat{T} \,|\, \mathbf{X}, \mathbf{Y}) &= \text{var}(\hat{T} \,|\, \mathbf{X}, \mathbf{Y}) + E^2(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) \\ &= \sum_{I=1}^{N} \sum_{J=1}^{N} \left\{ \frac{\text{cov}(S_I, S_J \,|\, \mathbf{X})}{\hat{E}(S_I \,|\, \mathbf{X}) \hat{E}(S_J \,|\, \mathbf{X})} + \left(\frac{E(S_I \,|\, \mathbf{X})}{\hat{E}(S_I \,|\, \mathbf{X})} - 1 \right) \left(\frac{E(S_J \,|\, \mathbf{X})}{\hat{E}(S_J \,|\, \mathbf{X})} - 1 \right) \right\} Y_I Y_J \\ &= \sum_{I=1}^{N} \sum_{J=1}^{N} \left[\frac{E\left\{ \left(S_I - \hat{E}(S_I \,|\, \mathbf{X}) \right) \left(S_J - \hat{E}(S_J \,|\, \mathbf{X}) \right) \,|\, \mathbf{X} \right\}}{\hat{E}(S_I \,|\, \mathbf{X}) \hat{E}(S_J \,|\, \mathbf{X})} \right] Y_I Y_J \,. \end{aligned} \tag{8}$$

It is straightforward to show that if $\hat{E}$ actually recovers $E(S_I \,|\, \mathbf{X})$, then (8) reduces to the earlier expression (4) for the design variance of $\hat{T}$.

We do not explore specification of $\hat{E}$ in this paper, beyond noting that many standard weighting methods, including poststratification and ratio estimation, are special cases. For example, ratio estimation can be characterised as replacing $E(S_I \,|\, \mathbf{X})$ by the weighted average

$$\hat{E}(S_I \,|\, \mathbf{X}) = \frac{\sum_{J=1}^{N} X_J S_J}{\sum_{J=1}^{N} X_J} .$$

Irrespective of whether the W_I are defined via (2) or via (5), the "classic" design based approach to choosing an optimal sampling strategy is to choose an appropriate distribution for **S** in order to make the mean squared error (8) above as small as possible, subject usually to a restriction on the sample size, or more generally, on the sum of the components of **S**.

Where no restriction is placed on **Y**, this is an impossible task - a result first noted by Godambe (1955). A short proof of this famous "nonexistence" result, essentially based on Basu (1971) goes as follows: Consider the population defined by $Y_1 > 0$ and $Y_J = 0$, $J \neq 1$. In this case (4) is zero (and so our strategy is efficient) if we select our sample so that $pr(S_1 = 1|\mathbf{X}) = 1$, so $E(S_1|\mathbf{X}) = 1$, and use the weighting scheme (2). In particular, this strategy remains efficient when we impose the further restriction $pr(S_2 = 1|\mathbf{X}) = 0$. However, this restricted strategy is no longer optimal if we apply it to another population where $Y_2 > 0$ and $Y_J = 0$, $J \neq 2$. Consequently, no "globally" optimal sampling strategy exists under the design based approach. Each sampling strategy needs to be looked at anew, since there is no "gold standard" against which it can be compared.

4. Optimality under the Model Assisted Approach

As the preceding paragraph makes clear, the main problem with the design based approach to finding an optimal sampling strategy is that it is far too general. By specifying efficiency criteria in terms of the conditional distribution of $\hat{T}$ given **X** and **Y**, this approach paints itself into a corner. As a consequence, almost from the very beginning of large scale application of design based theory in survey sampling, practitioners have adopted "design-unbiasedness" strategies which have "small design mean squared error" for those realisations of **Y** which are, in some sense, "reasonable".

In practice, such values of **Y** are typically defined by assuming a model for the distribution of **Y** given **X**. That is, practitioners have been willing to use models in order to identify "optimal" strategies for estimating T. However, their assessment of these strategies has remained design based.

4.1 Model assisted strategies that are also design unbiased

This model assisted approach is comprehensively discussed in the recent text by Särndal, Swensson and Wretman (1992). Typically, the approach still assumes that the weighting variable W at least approximately satisfies (2). That is, the resulting estimator $\hat{T}$ is design-unbiased, or approximately so. However, rather than attempting to specify the

distribution of the sample design variable S by minimising the design mean squared error (8) for all possible values of **Y**, the model assisted approach seeks to minimise the expectation of this quantity given **X**. That is, we seek a distribution for **S** which minimises the average value of (8) over those Y values "consistent with" the known values in **X**.

From (8) we see that this expected or "average" (design) mean squared can be written

$$E\left[MSE(\hat{T}\,|\,\mathbf{X},\mathbf{Y})\,|\,\mathbf{X}\right] = \sum_{I=1}^{N}\sum_{J=1}^{N}\left[\frac{E\left\{\left(S_I - \hat{E}(S_I\,|\,\mathbf{X})\right)\left(S_J - \hat{E}(S_J\,|\,\mathbf{X})\right)\,|\,\mathbf{X}\right\}}{\hat{E}(S_I\,|\,\mathbf{X})\hat{E}(S_J\,|\,\mathbf{X})}\right]E(Y_I Y_J\,|\,\mathbf{X})\ . \tag{9}$$

Given a specification for the first and second order moments of **Y** given **X**, this expression can be minimised, and an optimal sample design (and hence optimal sampling strategy) obtained. To illustrate, consider the case where the X_I are strictly positive and the regression of Y on X is linear and through the origin. That is

$$\begin{aligned} E(Y_I\,|\,\mathbf{X}) &= \beta X_I \\ \mathrm{var}(Y_I\,|\,\mathbf{X}) &= \sigma_I^2 \\ \mathrm{cov}(Y_I, Y_J\,|\,\mathbf{X}) &= 0;\ I \neq J. \end{aligned} \tag{10}$$

Then, when the weights W_I are determined by (2),

$$E\left[MSE(\hat{T}\,|\,\mathbf{X},\mathbf{Y})\,|\,\mathbf{X}\right] = \sum_{I=1}^{N}\left[\frac{\mathrm{var}(S_I\,|\,\mathbf{X})}{E^2(S_I\,|\,\mathbf{X})}\right]\sigma_I^2 + \beta^2\sum_{I=1}^{N}\sum_{J=1}^{N}\left[\frac{\mathrm{cov}(S_I, S_J\,|\,\mathbf{X})}{E(S_I\,|\,\mathbf{X})E(S_J\,|\,\mathbf{X})}\right]X_I X_J\ .$$

To make things even simpler, we restrict ourselves further to the case of Bernoulli sampling. Here S_I is either one or zero, with the I^{th} population unit either included into or excluded from the sample according to the outcome of an independent Bernoulli trial with "success" probability $\pi_I = E(S_I\,|\,\mathbf{X})$. Consequently

$$\mathrm{var}(S_I\,|\,\mathbf{X}) = E(S_I\,|\,\mathbf{X})\left[1 - E(S_I\,|\,\mathbf{X})\right] = \pi_I(1-\pi_I)$$

and, for $I \neq J$

$$\mathrm{cov}(S_I, S_J\,|\,\mathbf{X}) = 0.$$

The expected design mean squared error for this case is therefore

$$E\left[MSE(\hat{T}\,|\,\mathbf{X},\mathbf{Y})\,|\,\mathbf{X}\right] = \sum_{I=1}^{N}\left(\frac{1}{\pi_I} - 1\right)(\sigma_I^2 + \beta^2 X_I^2).$$

This expression is minimised, subject to the expected sample size constraint

$$\sum_{I=1}^{N} \pi_I = n$$

when

$$\pi_I = n \frac{\sqrt{\sigma_I^2 + \beta^2 X_I^2}}{\sum_{J=1}^{N} \sqrt{\sigma_J^2 + \beta^2 X_J^2}}. \qquad (11)$$

Observe that when σ_I is proportional to X_I this optimal sample inclusion probability is proportional to X_I. Furthermore, for $\sigma_I = \sigma >> \beta$, these optimal probabilities are approximately nN^{-1}. That is, in cases where the regression effect in (10) is insignificant, a strategy that has equal first order inclusion probabilities is indicated.

4.2 Model assisted strategies that are design unbiased on average

The requirement that $\hat{T}$ be design unbiased (or approximately design unbiased) that was imposed in 4.1 is rather strong. An appealing extension of the model assisted approach, whose motivation follows along the same lines as those leading to the use of the average mean squared error (9), is discussed in Brewer (1995). This replaces the design unbiasedness requirement by the weaker requirement that the design bias of $\hat{T}$ averages out to zero over possible values of Y. That is, rather than exact (or "approximately exact") design unbiasedness, one requires average design unbiasedness, or

$$E\left(\hat{T} - T \middle| \mathbf{X}\right) = E\left(E\left(\hat{T} - T \middle| \mathbf{X}, \mathbf{Y}\right) \middle| \mathbf{X}\right) = 0. \qquad (12)$$

Clearly, exact design unbiasedness implies average design unbiasedness. However, as we shall see in 5.1, there are many other design biased strategies which also satisfy (12). Consequently, this condition is rather weak. Under the regression model (10) for $\mathbf{Y}$, it translates as

$$E\left(E\left(\sum_{I=1}^{N}(W_I S_I - 1)Y_I \middle| \mathbf{Y}, \mathbf{X}\right) \middle| \mathbf{X}\right) = E\left(\sum_{I=1}^{N}(E(W_I S_I \middle| \mathbf{X}) - 1)Y_I \middle| \mathbf{X}\right)$$
$$= \beta \sum_{I=1}^{N}(E(W_I S_I \middle| \mathbf{X}) - 1)X_I = 0$$

or

$$\sum_{I=1}^{N} E(W_I S_I \middle| \mathbf{X}) X_I = \sum_{I=1}^{N} X_I .$$

There is no unique solution to this identity. In particular, all three of the following weighting methods satisfy it.

$$W_I = \frac{\sum_{J=1}^{N} X_J}{\sum_{J=1}^{N} E(S_J|\mathbf{X})X_J} \tag{13}$$

$$W_I = \frac{\sum_{J=1}^{N} X_J}{\sum_{J=1}^{N} S_J X_J} \tag{14}$$

and

$$W_I = \frac{\sum_{J=1}^{N} X_J}{E(S_I|\mathbf{X})\sum_{J=1}^{N} \frac{S_J X_J}{E(S_J|\mathbf{X})}}. \tag{15}$$

Note that the weights (13) and (14) do not depend on I. Furthermore (13) is the same for any realisation of **S**. In an asymptotic (large *N*, large *n*) sense, the weights defined by (13) and (14) are essentially the same, so an asymptotic analysis will lead to the same optimal sample design for both these weighting systems, provided one exists. Also, since

$$E\left[\sum_{J=1}^{N} \frac{S_J X_J}{E(S_J|\mathbf{X})}|\mathbf{X}\right] = \sum_{J=1}^{N} X_J$$

it follows that a similar asymptotic analysis indicates the weights defined by (15) and (2) are equivalent, so, in the case of the regression model (10) and Bernoulli sampling, the asymptotically optimal design under (15) is specified by the inclusion probabilities (11).

A large sample optimal sample design under either (13) or (14) can be obtained. As usual we assume the regression model (12) and Bernoulli sampling. Then

$$E\left(MSE\left(\hat{T}|\mathbf{Y},\mathbf{X}\right)|\mathbf{X}\right) = \sum_{I=1}^{N} \mathrm{var}\left(W_I S_I|\mathbf{X}\right)\left(\sigma_I^2 + \beta^2 X_I^2\right) + \sum_{I=1}^{N}\left\{E\left(W_I S_I|\mathbf{X}\right) - 1\right\}^2 \sigma_I^2$$

$$= \overline{\pi}^{-2}\left[\sum_{I=1}^{N} \pi_I(1-\pi_I)(\sigma_I^2 + \beta^2 X_I^2) + \sum_{I=1}^{N}(\pi_I - \overline{\pi})^2 \sigma_I^2\right]$$

where

$$\overline{\pi} = \left(\sum_{I=1}^{N} X_I\right)^{-1} \sum_{I=1}^{N} \pi_I X_I .$$

Provided σ_I / X_I does not increase with X_I, this average mean squared error is minimised by choosing a sample design that makes $\bar{\pi}$ as large as possible, subject to the usual sample size constraint. Such an optimal sample design is easily seen to be the "extreme" design that sets $\pi_I = 1$ for the n units in the population with largest values of X, and $\pi_I = 0$ for the remaining $N - n$ population units.

4.3 The robustness-efficiency tradeoff

If efficiency is the sole criterion for choice of a strategy and the assumed regression model (10) holds for the population being surveyed, then using the extreme sample design with weights defined by either (13) or (14) should lead to a much smaller average mean squared error than the more traditional design (11) with weights defined by either (2) or (15).

However, it is not unfair to say that most users of the model assisted approach would prefer the strategy specified by (11) and (15). In general, their main argument for this is that the weights (15), unlike (13) and (14), lead to an approximately design unbiased estimator (since they approximate the exactly design unbiased weights (2)), and consequently the sampling strategy specified by (11) and (15) is more "robust" to model misspecification than the strategy specified by the extreme design consisting of those n units with largest X-values, coupled with the weights specified by (13) or (14).

Since any model assumption is inevitably wrong, this argument, on the surface at least, seems reasonable. However, it is important to realise that the "robustness" achieved by design unbiasedness is a repeated sampling property. There is no guarantee that a sample generated via (11) and using the (approximately) design unbiased weights (15) will result in an estimate that is more <u>accurate</u> than the estimates obtained using the design biased weights (13) or (14) for the same sample. This issue is discussed in more detail in section 6.

5. Efficiency under the Model Based Approach

As the discussion in the previous two sections has made clear, the concept of design unbiasedness is crucial to both the design based as well as the model assisted approaches to defining a sampling strategy. However, under the model based approach this basic requirement

is abandoned. The argument for doing so is straightforward. Since the distributions of both **S** and **W** are completely known once **X** is known, their realisations contribute no information about **Y** (and functions of **Y**, like *T*) over and above that already provided by **X**. That is, once **X** is known, **S** and **W** become ancillary statistics for inference about **Y**. Consequently, application of the Conditionality Principle leads to the conclusion that any inference about *T* should be conditioned on **S** and **W**. Since design unbiasedness does not condition on these values (averaging in fact over all possible realisations of these statistics), it is an inappropriate criterion to apply to the estimator $\hat{T}$.

The model based approach has been most strongly linked with the work of Richard Royall and his students. An elegant summary of the philosophy behind this approach is set out in Royall (1976). Since design unbiasedness is no longer a requirement, the obvious alternative property we require of $\hat{T}$ under this approach is that it be model unbiased. That is,

$$E\left(\hat{T} - T \middle| \mathbf{S}, \mathbf{X}\right) = 0. \tag{16}$$

In other words, the values of the estimation errors $\hat{T} - T$ obtained for all population realisations **Y** consistent with the actual value of **X** observed, and the sample **S** actually obtained, should average out to zero. The natural measure of the accuracy of $\hat{T}$ as an estimator of T under this approach is then the variance of $\hat{T} - T$ given **S** and **X**.

In the context of the regression model (10) and the general linear estimator $\hat{T}$ the model unbiasedness condition (16) becomes

$$\sum_{I=1}^{N} (W_I S_I - 1) X_I = 0 \tag{17}$$

and an optimal sample design is then one that ensures the resulting conditional variance

$$\mathrm{var}\left(\hat{T} - T \middle| \mathbf{S}, \mathbf{X}\right) = \sum_{I=1}^{N} (W_I S_I - 1)^2 \sigma_I^2 = \sum_{I=1}^{N} \left\{ \Delta(S_I > 0)(W_I S_I - 1)^2 + \Delta(S_I = 0) \right\} \sigma_I^2 \tag{18}$$

is as small as possible. Here Δ denotes the indicator function that takes the value 1 if its argument is true and is zero otherwise.

Note that the variance criterion (18) does not depend on weights for nonsampled units. Consequently, these can be set to zero. The optimal weights for the sampled units are obtained by minimising (18) subject to (17). These turn out to be of the form

$$W_I = \frac{1}{S_I}\left\{\frac{X_I}{\sigma_I^2}\left[\frac{\sum_{J=1}^{N}\Delta(S_J=0)X_J}{\sum_{J=1}^{N}\Delta(S_J>0)\frac{X_J^2}{\sigma_J^2}}\right]+1\right\} \tag{19}$$

and it is straightforward to show that in this case

$$\hat{T} = \sum_s Y_I + \frac{\sum_s \frac{Y_I X_I}{\sigma_I^2}}{\sum_s \frac{X_I^2}{\sigma_I^2}} \sum_r X_I$$

which is the "classic" linear predictor of T under the model (10).

The final step in finding an optimal model based sample design is to identify the distribution for **S** that minimises (18) when **W** is defined by (19). Since in this case

$$\mathrm{var}\left(\hat{T}-T\middle|\mathbf{S},\mathbf{X}\right) = \frac{\left(\sum_{I=1}^{N}\Delta(S_I=0)X_I\right)^2}{\sum_{I=1}^{N}\Delta(S_I>0)\frac{X_I^2}{\sigma_I^2}} + \sum_{I=1}^{N}\Delta(S_I=0)\sigma_I^2$$

it is not difficult to see that, provided X_I / σ_I is nonincreasing in X_I, the optimal sample design is the extreme design (i.e. the one that selects the n units with largest X-values).

5.1 Model unbiased is also average design unbiased

It is interesting to observe that, for any set of model unbiased weights, (16) implies

$$E\left(\hat{T}-T\middle|\mathbf{X}\right) = E\left(E\left(\hat{T}-T\middle|\mathbf{S},\mathbf{W},\mathbf{X}\right)\middle|\mathbf{X}\right) = 0$$

and so all such weighting methods are also average design unbiased (12). Furthermore,

$$\begin{aligned} E\left[MSE(\hat{T}|\mathbf{X},\mathbf{Y})\middle|\mathbf{X}\right] &= E\left[E\left(\left(\hat{T}-T\right)^2\middle|\mathbf{X},\mathbf{Y}\right)\middle|\mathbf{X}\right] \\ &= E\left[\left(\hat{T}-T\right)^2\middle|\mathbf{X}\right] \\ &= E\left[E\left(\left(\hat{T}-T\right)^2\middle|\mathbf{S},\mathbf{W},\mathbf{X}\right)\middle|\mathbf{X}\right] \\ &= E\left[\mathrm{var}(\hat{T}-T|\mathbf{S},\mathbf{W},\mathbf{X})\middle|\mathbf{X}\right]. \end{aligned}$$

Since the conditional variance inside the square brackets in the last expression above is minimised whenever the extreme sample is chosen, it follows that the optimal model based design for $\hat{T}$ under (10) also minimises the average mean squared error of this estimator. That

is, the optimal model based design for $\hat{T}$ under (10) is also the optimal design for this estimator under the average design unbiased approach of section 4.2.

6. Choosing a Robust Strategy

So far we have concentrated on choice of an optimal strategy (if one exists) separately for each of the three approaches considered in this paper. In practice, however, one has to make a choice between these approaches for any particular application. How does one choose between the design based, model assisted and model based approaches to identifying a sampling strategy in such a case? One criterion that is often invoked in making such a choice is that of robustness. We choose the approach that leads to robust inference (i.e. inference that is somehow not strongly tied to assumptions about the conditional distribution of **Y** given **X**), and, within the chosen approach identify an optimal strategy.

Now, from the design based point of view, robustness is a non-issue, since inference under this approach does not need to model the conditional distribution of **Y** given **X**. Consequently, a naive user might argue that its nonparametric nature makes the design based approach the obvious methodology for choosing a sampling strategy. However, as we have seen earlier, this choice leads nowhere since there are no relevant optimality criteria that can be checked under this approach. If one wants both robustness and optimality, one must turn to the model assisted and model based approaches.

Both of these recognise that one has to model the distribution of a survey variable Y in terms of available frame information (X) in order to decide on a strategy. Where these two approaches diverge, however, is on the meaningfulness of imposing the requirement that the sampling strategy adopted be design unbiased (or at least approximately so). In particular, the model assisted approach claims robustness as a consequence of imposing (exact or approximate) design unbiasedness.

6.1 Robustness and design unbiasedness

The basic argument behind the imposition of design unbiasedness (exact or, more usually, approximate) is that of robustness. One allows the model to dictate the type of sample

selected, but one does not allow it to also dictate the type of weighting method used. The weights are typically constrained so that the repeated sampling average value of the estimator $\hat{T}$ is equal to, or is approximately equal to, the population total of Y no matter what model actually holds for Y in the population of interest. In the words of a colleague and staunch believer in the model assisted approach (Ken Brewer), adopting the model assisted approach is like wearing both a belt and braces to hold up one's trousers. If the belt (the model) should break, then one is not going to be totally embarrassed, since the braces (design unbiasedness) should still keep things in place.

From a model based point of view, however, this robustness argument is unconvincing. Since, as has already been pointed out, design unbiasedness is not a property associated with any particular sample, but rather one obtained by averaging over repeated samples, there does not appear to be any reason to believe that imposition of design unbiasedness on its own is sufficient to somehow protect the survey analyst from a large estimation error (due to model misspecification for example) in any particular sample. A very large positive error associated with sample 1 can be cancelled out by a corresponding large negative error associated with sample 2. "On average" things are fine, but for any particular sample they may be terrible. One has only to remember Basu's elephant fable (Basu, 1971) to realise how foolish complete reliance on design unbiasedness can be.

The standard counter argument to this criticism is that in large samples, the use of probability sampling methods allows the law of large numbers to be invoked, ensuring that a design unbiased (or approximately design unbiased) estimator will converge to the true value of T. Consequently the robustness property is really a large sample property. While this observation is certainly true, it also assumes that the survey analyst is only interested in large sample inference. It also fails to mention how large is "large". Central limit behaviour may require sample sizes considerably greater than the survey designer can afford. Furthermore, it leaves wide open the question of appropriate sample design for small to medium sample sizes. Many modern survey collections are run under very tight budgets, ruling out large sample sizes. Sample designs for these collections rely on design unbiasedness at their peril, and modern

model based designs are increasingly being used in these cases in an effort to maximise estimation efficiency.

6.2 Model based is not model dependent

As we saw in the preceding section, the model based approach can lead to extreme samples when taken to its logical conclusion. This has lead to strong criticism of the model based approach (Hansen, Madow and Tepping, 1983), since such extreme samples can lead to highly biased estimators if the model is misspecified. What this criticism ignores of course is that there is no particular reason why one should not investigate the sensitivity of an optimal model based design to breakdown of the model assumptions. Such analyses have in fact been a primary focus of model based strategies for some time, and they typically lead to non-extreme designs which are operationally very similar to many conventional designs.

To illustrate, the model based strategy defined by the extreme sample and the weights (19) becomes model biased if the true relationship between Y and X deviates from the strict proportionality relationship defined by (10). If the true relationship between Y and X is in fact described by a polynomial of degree K, say, then the optimal estimator defined by (10) remains unbiased provided the sample satisfies a K^{th} order "balance" condition, i.e. where the sample moments of X of order up to K equal their corresponding population moments (Royall and Herson, 1973). This particular "model robust" sample design is certainly NOT extreme.

It is important to realise that such model robust strategies are not the "blanket cures" claimed for probability sampling and design unbiasedness. They provide a reasonable level of efficiency over a chosen range of alternative potential models for the distribution of Y given X. In doing so, they lose efficiency at the assumed model (which generates the weights used in $\hat{T}$). This efficiency loss may be considerable if the range of potential alternative models is wide. In effect, the size of one's insurance premium goes up the greater the number of unpleasant events against which one wants to be protected. At the end of the day, it remains the survey designer's responsibility to carry out a sufficiently careful analysis of whatever data sources are available to ensure that the model underlying the chosen strategy is a good representation of the true distribution of Y given X in the population.

6.3 Robustness by adapting to the sample data

No amount of pre-selection analysis can prepare one for every eventuality. Models that seemed entirely appropriate before the sample data were obtained may suddenly look rather fragile when one has had a chance to actually check out the relationship between Y and X in the sample data. If one adopts a model based approach this situation is of no great concern. A crucial advantage to adopting this approach is its flexibility. There is no restriction that the model used to develop the sample selection procedures (the "design" model) should also be used in estimation.

In many cases there are distinct advantages in "widening" the scope of possible models for Y at the estimation stage of a survey, using the information collected in the sample. A common example of this is poststratification (Holt and Smith, 1979, Valliant 1993; Nascimento Silva and Skinner 1995). Another example is the widespread use of calibration weighting methods, where original sample weights derived at the time of sample selection (based perhaps on some preliminary "design" model for the population) are modified at the time of estimation so they result in estimates that are unbiased with respect to a final, more complex, "estimation" model for the population (Deville and Sarndal, 1992). A similar idea underlies the introduction of nonparametric adjustment factors based on a nonparametric smooth of the "design" model sample residuals (Chambers, Dorfman and Wehrly 1993). These adjustment factors can then be applied to the optimal weights under the "design" model to obtain final weights that are much more robust to model misspecification than the original optimal weights. Of course, these modified weights are no longer efficient under the "design" model, but, as always, one has to pay an efficiency premium for robustness.

As an interesting aside, it can be shown that these model based nonparametric weights are in some circumstances very similar to the (exactly or approximately) design unbiased weights derived under the model assisted approach (Chambers 1996). Consequently, there appears to be scope for these two apparently quite distinct approaches to lead to essentially the same sample inferences. Although further research is needed here, such a confluence may help resolve the sometimes overheated debate on which approach is "best".

6.4 Is probability sampling essential?

By probability sampling we mean a distribution for **S** such that $pr(S_I > 0 | \mathbf{X}) > 0$ for all *I*. This condition is an integral part of any sampling strategy under both the design based and model assisted approaches. This is because

(i) Efficiency is measured either by $\text{var}(\hat{T} - T | \mathbf{Y}, \mathbf{X})$ or by its expected value under the model, both of which are identically zero if the distribution of **S** is degenerate;

(ii) The requirement that $\hat{\text{T}}$ be exactly (or approximately) design unbiased leads to weights that satisfy (or approximately satisfy) (2). Consequently we require $E(S_I | \mathbf{X}) > 0$ for all *I*. This is guaranteed by probability sampling;

(iii) Robustness considerations under both approaches require application of the law of large numbers in order to guarantee that a design unbiased (or approximately design unbiased) estimator takes values arbitrarily close to the unknown population total T for large enough populations and samples.

In contrast, probability sampling is not essential under the model based approach. However, this does not mean that such sampling methods are excluded under this approach. Model based strategies are typically specified in terms of tight sample constraints (e.g. balance), but no restriction is placed on the actual method used to select the sample. There are good arguments (e.g. Royall 1976) for using a probability sampling method to actually select the final sample, subject to it satisfying these constraints, in order to avoid the unconscious bias that may creep into the selection process if a nonprobability method of sampling is employed.

In terms of the notation that has been used in this paper, this bias arises because the distributions of the population vectors **S** and **Y** are no longer independent given **X**. In particular, the distribution of **S** depends on **Y** as well as **X**. In such cases the model based (and model assisted) results presented in this paper are no longer valid. Alternative results can be derived, provided we can specify the nature of the dependence between **S**, **Y** and **X**. This is typically impossible, or at least very difficult. Consequently, a proponent of the model based approach will typically favour some form of probability sampling because this guarantees the distributions of **S** and **Y** are independent given **X**.

From a model based perspective, therefore, the principal argument for using probability sampling is to provide robustness against selection bias effects. However, there is another, more practical, aspect to probability sampling that makes it desirable from a model based point of view. This is the fact that "balanced" samples (in the general sense of balance, that is where the sample satisfies conditions which ensure unbiasedness of the proposed estimator within a specified class of alternative models that are "possible" for the survey population) are typically easier to achieve provided an appropriate form of randomisation is used.

For example, if the estimator of choice is the simple ratio estimator and the class of alternative models for the population is specified in terms of polynomial regression models of order up to and including K, then a balanced sample is one with all its X-moments up to and including order K equal to the corresponding population moments of X. On expectation over repeated sampling (i.e. in design expectation) a sample selected with equal inclusion probabilities for all population units will be balanced (Royall and Pfeffermann, 1982; Royall and Cumberland, 1988). Consequently, one way of achieving this type of balance is to take such a probability sample, and use it if its "balance" is adequate. Otherwise, we reject it, and select another probability sample. This idea of using probability-based "rejection" sampling to screen for adequately balanced samples has been shown to lead more precise inference than the corresponding use of unrestricted randomised sampling (Tam and Chan, 1984; Deville, 1992). In general, therefore, there is no tension between robust model based design and probability sampling. The former provides a criterion that the sample of choice should (at least approximately) satisfy, and the latter provides a mechanism for choosing samples to check against this criterion.

At the end of the day however, one has to ask oneself the question: Is there anything one can do if the underlying population model is such that our estimator, even when computed on a "balanced" sample, remains biased? That is, the real population model is not a member of the class of models underlying the chosen balance criteria, and so balance does not guarantee unbiasedness with respect to "reality". Does the fact that this sample has been selected via some form of randomisation based procedure help? Here it seems that one has no recourse but to design unbiasedness. That is, the only statements one can make relate to average properties

of the estimator over repeated sampling, rather than to the properties of the estimator for the actual sample selected. Since, as we have already pointed out, these average properties may be far from the actual behaviour of the estimator over the chosen sample, the inevitable conclusion one has to draw is that one cannot be protected against everything, and so one has to accept some risk in survey inference. The key property of good sample design is that it minimises this risk (by appropriate choice of model, balance criteria etc.) subject to available resources.

6.5 What is the right way to measure precision?

The astute reader will no doubt by now have asked the question: Efficiency of estimation is all very well, but the bottom line in any statistical analysis of sample survey data must be an accurate measure of the precision of that analysis. Where is the discussion on how to measure precision properly? Should one measure the precision of an estimator $\hat{T}$ by its design based error variance $\text{var}(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y})$ or should one measure it by its model based error variance $\text{var}(\hat{T} - T \,|\, \mathbf{S}, \mathbf{X})$?

Which measure is appropriate depends very much on <u>what</u> one means by precision and <u>when</u> one is measuring precision. Assuming unbiasedness of $\hat{T}$, we take precision as being the variance of the estimation error $\hat{T} - T$ with respect to all relevant sources of uncertainty at any particular point in time. Thus, one could argue that since **S** is unknown prior to sample selection, the design based error variance is a measure of our uncertainty about the estimation error <u>before</u> the sample is selected. However, it does condition on **Y**, which is also unknown before the sample is selected (and only partially known afterwards). Consequently, a better measure of precision before sampling is $\text{var}(\hat{T} - T \,|\, \mathbf{X})$.

<u>After</u> the sample is selected, however, and remembering that we are assuming the outcome **S** is ancillary (e.g. through probability sampling), it is clear that the appropriate measure of precision is at least the model based frequentist variance $\text{var}(\hat{T} - T \,|\, \mathbf{S}, \mathbf{X})$, or, if one is adopting a Bayesian approach, the posterior variance $\text{var}(\hat{T} - T \,|\, \mathbf{S}, \mathbf{X}, \mathbf{Y}_s)$, where $\mathbf{Y}_s$ is the vector of sample *Y*-values. There is strong empirical evidence (e.g. Royall and Cumberland, 1981) that a variance which does not condition on **S** (like the design based variance) can give a very misleading picture of the precision of $\hat{T}$ once **S** is <u>known</u>.

The situation gets even more complicated when we consider the problem of estimating precision. There are well known methods for estimating the design based error variance $\text{var}(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y})$ (see Wolter, 1985). Such estimators have model based properties as well however. Let $\hat{V}$ denote a design unbiased estimator of $\text{var}(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y})$. That is $E(\hat{V} \,|\, \mathbf{X}, \mathbf{Y}) = \text{var}(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y})$. Suppose also (as will usually be the case under either the model assisted or model based approaches) that $\hat{T}$ is model unbiased. Then

$$\begin{aligned}
\text{var}(\hat{T} - T \,|\, \mathbf{X}) &= E\Big(\text{var}(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) \,\Big|\, \mathbf{X}\Big) + \text{var}\Big(E(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) \,\Big|\, \mathbf{X}\Big) \\
&= E\Big(E(\hat{V} \,|\, \mathbf{X}, \mathbf{Y}) \,\Big|\, \mathbf{X}\Big) + \text{var}\Big(E(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) \,\Big|\, \mathbf{X}\Big) \\
&= E(\hat{V} \,|\, \mathbf{X}) + \text{var}\Big(E(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) \,\Big|\, \mathbf{X}\Big) \\
&= E\Big(E(\hat{V} \,|\, \mathbf{S}, \mathbf{X}) \,\Big|\, \mathbf{X}\Big) + \text{var}\Big(E(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) \,\Big|\, \mathbf{X}\Big)
\end{aligned}$$

while

$$\begin{aligned}
\text{var}(\hat{T} - T \,|\, \mathbf{X}) &= E\Big(\text{var}(\hat{T} - T \,|\, \mathbf{S}, \mathbf{X}) \,\Big|\, \mathbf{X}\Big) + \text{var}\Big(E(\hat{T} - T \,|\, \mathbf{S}, \mathbf{X}) \,\Big|\, \mathbf{X}\Big) \\
&= E\Big(\text{var}(\hat{T} - T \,|\, \mathbf{S}, \mathbf{X}) \,\Big|\, \mathbf{X}\Big)
\end{aligned}$$

so

$$E\Big(E(\hat{V} \,|\, \mathbf{S}, \mathbf{X}) - \text{var}(\hat{T} - T \,|\, \mathbf{S}, \mathbf{X}) \,\Big|\, \mathbf{X}\Big) = -\text{var}\Big(E(\hat{T} - T \,|\, \mathbf{X}, \mathbf{Y}) \,\Big|\, \mathbf{X}\Big)$$

That is, the average model bias of the design unbiased variance estimator $\hat{V}$ is equal to minus the average variance of the design bias of the estimator $\hat{T}$. In general, therefore, the design unbiased variance estimator $\hat{V}$ will be biased low for the "correct" post-sample uncertainty of the estimator $\hat{T}$. One situation where $\hat{V}$ will be a reasonable measure of this uncertainty is where the sample design ensures that the average design bias of $\hat{T}$ varies little between different realisations of *Y*. This will be the case if $\hat{T}$ is also design unbiased, or approximately design unbiased. Sample designs that ensure this condition is satisfied are typically those that lead to "balanced" samples for $\hat{T}$. Consequently, design based variance estimators like $\hat{V}$ are usually "safe" (in the sense of actually estimating the right thing) in balanced samples. In unbalanced samples, however, they are not to be trusted.

Of course, model unbiased variance estimators can also be derived, and these will provide "correct" measures of precision irrespective of the type of sample selected. However, these variance estimators will no longer be correct if the assumed model is misspecified. Hence

robustness of variance estimation is as important as robustness of estimation under the model based approach.

In a series of papers, Royall and Eberhardt (1975) and Royall and Cumberland (1978, 1981a, 1981b, 1985) have explored a general approach to robustifying standard least squares type model based variance estimators. Their method assumes correct specification of the conditional mean of Y given X and uses a nonparametric moment estimator (rather than a parametric one) for the leading term in the conditional variance $\text{var}(\hat{T} - T \,|\, \mathbf{S},\mathbf{X})$. Empirical results presented by these authors indicate that the general performance of this robust approach to variance estimation (including confidence interval uniformly good provided samples are balanced, or are close to balance. In unbalanced samples, however, presence of bias in the estimator $\hat{T}$ can lead to substantial noncoverage.

Consequently, at the time of writing there does not appear to be a general solution (either design based or model based) to estimating the precision of a sample survey estimator a elected. In particular, accurate variance and confidence interval estimation in unbalanced samples remains an area of current research.

Acknowledgements

The original version of this paper was presented as part of the Workshop on Survey Design and Analysis at the Australian Statistical Congress, Sydney, July 8 - 12 1996. Danny Pfeffermann, Alan Dorfman, Fred Smith, Phil Kokic, Roger Sugden and Chris Skinner all provided comments and suggestions on a subsequently revised version of the paper. Their valuable inputs are gratefully acknowledged. All remaining errors and omissions are the author's sole responsibility, however.

References

BASU, D. (1971). An essay on the logical foundations of survey sampling I. In *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

BREWER , K. R. W. (1995). Combining design based and model based inference. Chapter 30 in *Business Survey Methods* (editors: Cox, Binder, Chinnappa, Christianson, Colledge and Kott). New York: John Wiley.

BREWER, K. R. W. & SÄRNDAL, C. E. (1983). Six approaches to enumerative survey sampling. In *Incomplete Data in Survey Sampling 3*, Session VIII, 363-368, New York: Academic Press.

CHAMBERS, R. L. (1996). Robust case weighting for multipurpose establishment surveys. To appear in *Journal of Official Statistics*.

CHAMBERS, R. L., DORFMAN, A. H. AND WEHRLY, T. E. (1993). Bias robust estimation using nonparametric calibration. *Journal of the American Statistical Association* **88**, 268 - 277.

COCHRAN, W. G. (1977). *Sampling techniques*. New York: John Wiley.

DEVILLE, J. C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, Statistics Sweden, Örebro, October 5 - 7 1992.

GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society* **B 17**, 269 - 278.

HANSEN, M. H., MADOW, W. G. & TEPPING, B. J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association* **78**, 776 - 807.

HOLT, D. & SMITH, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society* **A 142**, 33 - 46.

KISH. L. (1965). *Survey sampling*. New York: John Wiley.

NASCIMENTO SILVA, P. L. D. AND SKINNER, C. J. (1995). Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics* **11**, 277 - 294.

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 - 625.

PFEFFERMANN, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review* **61**, 317 - 337.

ROYALL, R. M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology* **104**, 463 - 474.

ROYALL, R. M. & HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association* **68**, 880 - 889.

ROYALL, R. M. & EBERHARDT, K. A. (1975). Variance estimates for the ratio estimator. *Sankhya* **C 37**, 43 - 52.

ROYALL, R. M. & CUMBERLAND, W. G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association* **73**, 351 - 358.

ROYALL, R. M. & CUMBERLAND, W. G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association* **76**, 66 - 88.

ROYALL, R. M. & CUMBERLAND, W. G. (1981b). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society* **B 43**, 353 - 367.

ROYALL, R. M. & CUMBERLAND, W. G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association* **80**, 355 - 359.

ROYALL, R. M. & CUMBERLAND, W. G. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society* **B 50**, 118 - 124.

SÄRNDAL, C.-E., SWENSSON, B. AND WRETMAN, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

SMITH, T. M. F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society* **A 139**, 183 - 204.

SMITH, T. M. F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society* **A 146**, 394 - 403.

SMITH, T. M. F. (1984). Sample surveys: Present position and potential developments: Some personal views. *Journal of the Royal Statistical Society* **A 147**, 208 - 221.

SMITH, T. M. F. (1994). Sample surveys 1975 - 1990; An age of reconciliation? *International Statistical Review* **62**, 3 - 34.

SUGDEN, R. A. & SMITH, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495 - 506.

TAM, S. M. & CHAN, N. N. (1984). Screening of probability samples. *International Statistical Review* **52**, 301-308.

VALLIANT, R. (1993).Poststratification and conditional variance estimation. *Journal of the American Statistical Association* **88**, 89 - 96.

WOLTER, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.