DACSEIS

IST-2000-26057

Workpackage 11

Imputation and Non-response

Deliverable 11.1

2004

List of contributors:

Yves G. Berger, University of Southampton
Jan Bjørnstad and Li-Chun Zhang, Statistics Norway
Chris Skinner, University of Southampton

Main responsibility:

Yves G. Berger, University of Southampton
Jan Bjørnstad and Li-Chun Zhang, Statistics Norway
Chris Skinner, University of Southampton

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

Preface

This document discusses aspects of variance estimation for complex survey data in the presence of non-response. Chapter 1 provides an overview of the problem and of the underlying statistical framework. Chapter 2 discusses the nature of the variance to be estimated, outlining a number of alternative approaches. The impact of non-response on the variance is discussed in Chapter 3, where a new measure of variance inflation, the *neff*, is introduced. Chapter 4 provides the main discussion of the central topic of this document: variance estimation. The chapter provides a review of existing methods but also introduces two new approaches: a general jackknife method for data subject to impution in Section 4.3 and a non-Bayesian version of multiple imputation in Section 4.4. The primary focus is on variance estimation for data subject to imputation, but the final Section of the document, 4.5, provides a brief review of variance estimation for weighted estimators.

Yves G. Berger — University of Southampton
Jan Bjørnstad and Li-Chun Zhang — Statistics Norway
Chris Skinner — University of Southampton

Contents

Chapter 1

Introduction

1.1 Point estimation for complex survey data in the presence of non-response

Sample surveys are subject to both *unit non-response* and *item non-response.* Unit non-response arises when no survey data are collected for a unit. Item non-response arises when some data are collected for a unit but values of some items are missing. Some patterns of non-response are more complex and might be viewed as examples of either unit or item non-response. See Groves *et al.* (2002) for further discussion of the sources and reasons for non-response in surveys.

Different approaches to point estimation may be adopted in the presence of non-response. Some methods just ignore the non-response. In the case of unit non-response this will usually involve treating the set of responding units as if it were the selected sample. In the case of item non-response this may involve deleting units which have missing values on any of the variables used in a particular analysis (*available cases analysis*) or deleting units which have missing values on any of the survey variables (*complete cases analysis*). Such approaches may be subject to bias and, in general, do not make most efficient use of the data. We shall not consider them explicitly, although they will typically feature as special cases of the approaches we do consider.

Weighting and *imputation* are the two main methods used to correct for bias due to non-response and to make efficient use of data. Weighting is classically used to treat the problem of unit non-response, whereas imputation is classically used to treat problems of item non-response.

Weighting is a 'unit-level' adjustment, providing a common form of adjustment for all analyses based a common set of responding units and is thus natural for the treatment of unit non-response. It is less practical to use weighting to treat item non-response, since a different method of weighting would be required for estimates based upon different sets of variables.

In contrast to weighting, imputation is a variable-specific adjustment and is thus natural to treat missing data in a given variable. Imputation tends to become more complicated

and time consuming to implement the more the variables treated and thus it is not usually considered as a practical solution for unit non-response in a survey measuring many variables.

1.2 Properties of point estimators in the presence of non-response

1.2.1 Bias and variance

In a classical frequentist framework for statistical inference, it is usual to summarise the properties of a point estimator in a sample survey in terms of its bias and variance. It is usually supposed that the sampling distribution of the estimator is approximately normal so that the characteristics of this distribution can be captured just by these two characteristics. For a survey employing a probability sampling scheme for which there are no non-sampling errors, it is usually possible to construct point estimators of standard population parameters of interest, which are approximately unbiased. Alternative approximately unbiased point estimators may then be compared according to their variances.

The presence of non-response will usually introduce bias into point estimators. A primary purpose of weighting and imputation methods is to reduce this bias but, in practice, it is unlikely that these methods will remove bias entirely. In addition to affecting bias, non-response will also affect the variance of a point estimator. The focus of this document will be on variance rather than bias and the impact of non-response on variance is investigated in Chapters 2 and 3.

In the absence of non-sampling errors, the classical design-based approach in sample surveys, following NEYMAN (1934), is to evaluate the bias and variance of the point estimator with respect to randomised sampling scheme. When non-response arises it is common to evaluate the bias and variance with respect to both the sampling scheme and the *non-response mechanism*. This mechanism is often represented probabilistically, like a probability sampling mechanism. A key difference between these two mechanisms is, however, that the sampling mechanism should generally be known (at least up to essential features of the mechanism), whereas the non-response mechanism will typically be unknown.

1.2.2 Non-response mechanisms

In the mainstream statistical literature, non-response is treated as a form of *missing data*. (There are other sources of missing data. For example, in a two-phase survey measurements for variables recorded in the second phase of the survey will be missing by design for units present in the first phase sample but not in the second phase sample.) It is therefore common to refer to non-response mechanisms as *missing data mechanisms*. We shall refer to the following types of mechanism. These are discussed in more detail in deliverable D11.2 and in LITTLE and RUBIN (2002).

Missing Completely At Random (MCAR): The values of the set of variables used to construct a point estimator are missing completely at random if missingess is independent of all these variables.

Missing at Random (MAR): The values of the set of variables used to construct a point estimator are missing at random given an additional set of measured variables if missingess is independent of the values of the variables which are missing, conditional on the observed values of both sets of variables.

Ignorable missingness: The missing data mechanism is ignorable if the properties of a given inference procedure are not affected by the nature of the mechanism.

Non-ignorable missingness: The missing data mechanism is non-ignorable if it is not ignorable.

1.3 Variance estimation

Since non-response will generally affect the variance of a point estimator, as will be discussed in Chapters 2 and 3, it is important to consider how to estimate this variance in the presence of non-response. Chapter 4 reviews existing methods and introduces some new methods for variance estimation with an emphasis on the estimation of variances for point estimators based upon imputed data. The first three sections of this chapter treat the case of single imputation, where each missing value has been replaced by just a single imputed value. The fourth section considers the method of multiple imputation. The final section provides some brief remarks about variance estimation for point estimators which use weighting to compensate for non-response.

Chapter 2

Variances in the presence of non-response

2.1 Introduction

Before considering variance estimation it is important to be clear about the nature of the variance being estimated. In this chapter we discuss the definition of variances of point estimators in the presence of non-response. We shall focus on point estimators including imputed data although much of the discussion is also relevant to weighted estimators and we refer to this case briefly in Section 2.6.

2.2 Framework

In this section we introduce the basic survey set-up to be considered. We suppose that the aim is to estimate a parameter θ, which is a function of the values of a $k \times 1$ vector of variables y_i for units $i = 1, \ldots, N$ in a finite population U. For example, we may be interested in a population total $\sum_U z_i$, where z_i is a component of y_i. We assume for simplicity that θ is a scalar. We suppose that a survey is undertaken on a sample, s, of U, selected by a probability sampling design, with the aim of measuring y_i for each $i \in s$. Because of non-response, however, we suppose that each variable in y_i may only be observed for a subset of units in s. Writing $y_i = (y_{1i}, \ldots, y_{ki})'$, we let $R_i = (R_{1i}, \ldots, R_{ki})'$ where $R_{ji} = 1$ if y_{ji} is observed (without error) and $R_{ji} = 0$ otherwise, $j = 1, \ldots, k, i \in s$.

In the classical approach to imputation a single imputed value (or vector of values) y^*_{ji} is created for each case (i, j) where $R_{ji} = 0$ and point estimation of θ takes place as if $y^*_{ji} = y_{ji}$. Thus let $Y^{(s)}$ be the $k \times n$ matrix with columns $y_i, i \in s$ and suppose $\widehat{\theta}(Y^{(s)})$ would be the estimator of θ if $Y^{(s)}$ were fully observed. For example, if $\theta = \sum_1^N Y_i$ then $\widehat{\theta}$ might take the form $\widehat{\theta} = \sum_s w_i y_i$ where the w_i are survey weights. Let $\tilde{y}_{ji} = y_{ji}$ if $R_{ji} = 1$ and $\tilde{y}_{ji} = y^*_{ji}$ if $R_{ji} = 0$, let $\tilde{y}_i = (\tilde{y}_{1i}, \ldots, \tilde{y}_{ki})'$ and let $\tilde{Y}^{(s)}$ be the $k \times n$ matrix with columns $\tilde{y}_i, i \in s$. Then $\widehat{\theta}(\tilde{Y}^{(s)})$ is the point estimator of θ which treats the imputed values y^*_{ji} as if they are the actual values y_{ji}.

The classical approach may be modified in a number of ways. First, the approach may be modified by estimating θ by a different function of $\tilde{Y}^{(s)}$, to correct for bias effects of imputation, (see for example, SKINNER and RAO, 2002). Second, repeated imputed values $y_{ji}^{*(1)}, \ldots, y_{ji}^{*(M)}$ may be created for each case where $R_{ji} = 0$, as in the methods of multiple imputation (Chapter 4) and fractional imputation (FAY, 1996). Let $\widehat{\theta}^*$ denote the resulting estimator of θ, both under these alternative approaches and in the classical case when $\widehat{\theta}^* = \widehat{\theta}(\tilde{Y}^{(s)})$.

We are interested in the sampling distribution of $\widehat{\theta}^* - \theta$ and in the next section will discuss alternative sources of stochastic variation with respect to which this distribution may be defined. This paper is concerned with estimating the variance $\mathrm{V}^* = \mathrm{var}(\widehat{\theta}^* - \theta)$ of this distribution. We shall generally only be concerned with situations where $\widehat{\theta}^*$ is approximately unbiased for θ and the sampling distribution of $\widehat{\theta}^* - \theta$ is approximately normal, $N(0, \mathrm{V}^*)$, so that V^* is a useful single summary measure of the accuracy of $\widehat{\theta}^*$ as an estimator of θ.

2.3 Sources of stochastic variation

There are four basic sources of stochastic variation with respect to which the sampling distribution of $\widehat{\theta}^*$ may be defined:

ξ : a *model* generating the population values $y_1, \ldots, y_N$;

p : the *sampling mechanism* used to select s from U.

q : the *response mechanism* generating $R_i, i \in s$;

I : the *imputation mechanism*; which may involve a stochastic element, for example in the selection of donors in a hot deck method or in the addition of noise in a regression imputation method.

The variance may be defined with respect to different combinations of these sources of variation, as will be discussed in Section 4.

Before considering these possibilities we introduce one extension of our framework. A number of authors extend the definition of R_i to all $i \in U$ (e.g. RUBIN, 1987, p. 30). The random variables R_i for $i \in \overline{s} = U \backslash s$ (the non-sampled units in U) may be constructed in an arbitrary way provided

$$p\left(R_s|s\right) = \int p\left(R_U|s\right) dR_{\overline{s}}$$

remains the response mechanism q, where $\mathrm{V}_{q^U p}(\widehat{\theta})$ and R_U denote the matrices with columns R_i for $i \in s, i \in \overline{s}$ and $i \in U$ respectively. Here we use $p(.)$ to denote a generic probability mass function. Many authors assume that R_U may be constructed so that the following assumption holds.

A1. **Assumption of independence of sampling and non response**: s and R_U are independent.

Under this assumption, the distribution of the $R_i, i \in U$ may be interpreted as a response mechanism which holds whether or not the i are included in the sample (e.g. Rubin, 1987, p. 30). We denote the distribution $p(R_U)$ as q^U and refer to it as the *population response mechanism*, when the above assumption holds. A consequence of this assumption is that $p(s|R_U) = p(s)$, that is that the distribution of s is the same whether or not we condition on R_U. This is the basis of a number of approaches. One is to condition on R_U and (when $k = 1$) to treat cases with $R_i = 0$ or 1 as respondent and non-respondent strata (e.g. Cochran, 1977, Section 13.2). Another is to evaluate the variance with respect to $p(s) = p(s|R_U)$ first and then with respect to $p(R_U)$ (e.g. Shao and Steel, 1999).

Assumption A1 seems a plausible one in many applications in practice but it is not without force. For example, a consequence of the assumption is that $p(R_i|s) = p(R_i)$ for $i \in s$, i.e. given that a unit is included in the sample, the probability that the unit will respond does not depend on which other units are also included in the sample. Note, however, that no assumption is implied about the relation between the R_i, for example it is *not* implied that R_i and R_j must be independent for $i \neq j$. See Fay (1991) and Lee *et al.* (2002) for further discussion of the framework implied by this assumption.

2.4 Definition of variance

In this section we consider possible candidates for the definition of the variance with respect to one or more of the sources of stochastic variation described in Section 2.3. These candidates cover the usual definitions employed in the literature.

Design-based Variance (V_p or V_{pI})

Consider first a deterministic approach where each y^*_{ji} is a given function of observed values of y_{ji} and other known sample or population information. For example, under ratio imputation with a single variable ($k = 1$)

$$y^*_i = x_i \overline{y}^{(s)}_{obs} / \overline{x}^{(s)}_{obs}, \quad i \in s$$

where x_i is a variable known for all $i \in s$ and $\left(\overline{y}^{(s)}_{obs}, \overline{x}^{(s)}_{obs}\right)$ denotes the mean of (y_i, x_i) for $i \in s$ with $R_i = 1$. If in this case θ is the population mean of $y_i, \theta = \overline{Y} = N^{-1} \sum_U y_i$, and $\widehat{\theta}^* = n^{-1} \sum_s \tilde{y}_i$ then it follows that

$$\widehat{\theta}^* = \left(\overline{y}^{(s)}_{obs} / \overline{x}^{(s)}_{obs}\right) n^{-1} \sum\nolimits_s x_i \tag{2.1}$$

For a much larger class of imputation methods and estimators, $\widehat{\theta}^*$ may be expressed as a known function of means.

$$\widehat{\theta}^* = h\left(n^{-1} \sum_s z_i\right) , \tag{2.2}$$

where z_i is a vector of elements, well-defined for all $i \in U$ and not dependent upon s, which will typically either be of the form $w_i R_i y_{ji}$, where w_i is a fixed survey weight, or of the form $w_i x_i$ for an auxiliary variable observed for all $i \in s$. In this case $V_p(\widehat{\theta}^*)$ is

well-defined and provides one possible variance measure, where the R_{ji} are fixed for $i \in U$ and may be interpreted as defining domains or 'non-response strata' (COCHRAN, 1977, Section 13.2) with respect to which the totals $\sum z_i$ are defined.

The measure V_p has the usual advantage of design-based variances, that it is not model dependent, but it does require assumption A1 so that R_U can be held fixed and hence is not entirely free of assumptions about the response mechanism.

The measure V_p allows for the variance inflation arising from the reduction in the sample size due to non-response, just as the design variance of a domain mean allows for the fact that the domain sample size will usually be smaller than the full sample size. It fails, however, to capture any element of the uncertainty about the difference between respondents and non-respondents. Consider, for example, the simpler version of (2.1) where $k = 1$, $x_i = 1$ and $\widehat{\theta}^*$ is simply the respondent mean $\widehat{\theta}^* = \overline{y}_{obs}^{(s)}$. Under simple random sampling the expectation of $\widehat{\theta}^*$ is $\overline{Y}_{obs}^{(U)} = \sum_U y_i R_i / \sum_U R_i$. The variance $\mathrm{V}_p(\widehat{\theta}^*)$ fails to reflect any difference between $\overline{Y}_{obs}^{(U)}$ and $\overline{Y}$. Even if the response mechanism is uniform, that is if all the R_i have a common Bernoulli distribution, then $\overline{Y}_{obs}^{(U)}$ and $\overline{Y}$ will in general not be identical. There is thus a (conditional) bias. This conditional bias can be captured in the uncertainty measure if the R_i are not conditioned upon and this provides motivation for the approach in the next section.

When the imputation approach is stochastic, it is natural to replace $\mathrm{V}_p(\widehat{\theta}^*)$ by:

$$\mathrm{V}_{pI}\left(\widehat{\theta}^*\right) = \mathrm{V}_p\left[\mathrm{E}_I\left(\widehat{\theta}^*\right)\right] + \mathrm{E}_p\left[\mathrm{V}_I\left(\widehat{\theta}^*\right)\right] \tag{2.3}$$

As in the deterministic case, this measure will fail to capture the estimation error (conditional bias) represented by the difference between $\overline{Y}_{obs}^{(U)}$ and $\overline{Y}$ when $k = 1$.

Quasi Design-based Variance with Population Response Mechanism $\mathrm{V}_{q^U p}$ (or $\mathrm{V}_{q^U pI}$)

The pure design variance V_p in the previous section failed to take account of estimation error arising from the difference between the responding and nonresponding parts of the population. This error may be accounted for by evaluating the variance not only with respect to the sample design p but also with respect to the population response mechanism q^U. As in the previous section, this formulation depends upon assumption A1. For deterministic imputation the combined variance may be expressed as:

$$\mathrm{V}_{q^U p}\left(\widehat{\theta}^*\right) = \mathrm{E}_{q^U}\mathrm{V}_p\left(\widehat{\theta}^*\right) + \mathrm{V}_{q^U}\mathrm{E}_p\left(\widehat{\theta}^*\right), \tag{2.4}$$

and for stochastic imputation

$$\mathrm{V}_{q^U pI}\left(\widehat{\theta}^*\right) = \mathrm{E}_{q^U}\mathrm{V}_{pI}\left(\widehat{\theta}^*\right) + \mathrm{V}_{q^U}\mathrm{E}_{pI}\left(\widehat{\theta}^*\right). \tag{2.5}$$

These are the variances considered by SHAO and STEEL (1999). Returning to the simple example in the previous section where $k = 1$, $\widehat{\theta}^* = \overline{y}_{obs}^{(s)}$ is the respondent mean and when the design is such that $\mathrm{E}_p(\widehat{\theta}^*) = \overline{Y}_{obs}^{(s)}$, it is now feasible that $\widehat{\theta}^*$ is unbiased for θ under the distribution with respect to both q^U and p, provided q^U obeys strong conditions.

Quasi Design-based 'Two-phase' Variance V_{pq} (or V_{pqI})

The approach in the previous section required the construction of q^U and an assumption of the form A1. An alternative and seemingly more natural approach is to treat the response mechanism as if it were the second phase of a sampling scheme, where p defines the first phase.

This approach is equivalent to that in the previous section if Assumption A1 holds. For, let $H\left(s, R_s, y_{obs}\right)$ be any statistic, where y_{obs} is the vector of values of y_i for which $R_i = 1$ and $i \in s$, then

$$\begin{aligned}\mathrm{E}_{q^U p} H\left(s, R_s, y_{obs}\right) &= \sum_{R_U} \sum_{s} H\left(s, R_s, y_s\right) p\left(s|R_U\right) p\left(R_U\right) \\ &= \sum_{s} \sum_{R_U} H\left(s, R_s, y_s\right) p(s) p\left(R_U\right)\end{aligned}$$

if Assumption A1 holds

$$\begin{aligned}&= \sum_{s} \sum_{R_s} H\left(s, R_s, y_s\right) p(s) p\left(R_s\right) \\ &= \mathrm{E}_p \mathrm{E}_q H\left(s, R_s, y_s\right)\end{aligned}$$

and the moments of any statistic are the same under either representation.

In principle, this approach allows the specification of the response mechanism to depend upon s. In practice, however, it is usual to treat q as if it were equivalent to a population response mechanism q^U restricted to s. For example, Rao and Shao (1992) assume a uniform Bernoulli response mechanism within specified imputation classes.

The two-phase variance V_{pq} extends naturally to V_{pqI} in the case of stochastic imputation (e.g. Rao and Shao, 1992).

Model-anticipated Variance $\mathrm{E}_\xi \mathrm{V}_{pq}$ (or $\mathrm{E}_\xi \mathrm{V}_{pqI}$)

A potential problem with the quasi design-based approaches in the previous two sections is that the response mechanism (q or q^U) is unknown and must be fully specified.

An alternative approach is to consider introducing dependence upon a model for the y_i into the variance.

While this adds model assumptions it may also reduce the assumptions required about the response mechanism. Deville and Särndal (1994) consider imputation based upon a model ξ for the relation between the y_i and auxiliary information for which it may be assumed that $\widehat{\theta}^*$ is model unbiased.

2.5 Relation to other DACSEIS Workpackages

In this section we consider how the alternative definitions of variance in Section 2.4 relate to the treatment of non-response in the other DACSEIS workpackages.

We first consider the main Monte-Carlo simulation study described in deliverables D3.1 and D3.2 The study first modifies the universe by non-response to obtain the observable universe. This observable universe is then sampled repeatedly to produce a simulation-based estimate of the variance. This corresponds to the design-based variance in Section 2.4. The process of modification of the universe corresponds to the population response mechanism q^U. The responding universe is held fixed in the simulation so the simulation-based variance corresponds to V_p (or to V_{pI} in the case of stochastic imputation).

The nature of this design-based variance is discussed further in Section 6.6 of deliverable D1.1 There, it is assumed that there exists a population response mechanism and the design-based variance is referred to as the conditional variance, since it conditions on the response indicators. It is compared with the quasi-design based variance (in the terminology of Section 2.4) which is referred to as the unconditional variance since it is with respect to both the sampling design and the response mechanism. The difference between the conditional and unconditional variances is explored and it is found, as discussed above, that there will be little difference between the two in practice if the sampling fraction is negligible, but that there may be an important difference if this is not the case.

Workpackage 5 on resampling methods considers imputation methods which allow for nonresponse. A number of variance estimation methods from the literature are reviewed. In addition, some variance estimators for deterministic imputation are developed. The basic approach is again design-based in the sense that the imputed estimator is represented as a sample statistic and its variance is estimated in the same way that the variance of a sample statistic would be estimated were there full response. The response mechanism is not explicitly allowed for in the variance estimator. The simulation study in deliverable D5.2 follows a two-phase approach where respondents are subsampled according to a uniform response mechanism. Hence the simulation variance is effectively a two-phase variance. Nevertheless, the sampling fractions within strata are very small so that, as noted in Section 6.6 of deliverable D1.1, the distinction between this two-phase variance and the design-based variance is negligible.

Workpackage 8 includes nonresponse in its simulation study when evaluating the variances of estimators. The variance estimators considered are essentially design-based again, treating the responding units as a fixed domain. As in workpackage 5, the simulation study follows a two-phase approach but again the sampling fractions are very small so the distinction between the two-phase variance and the design-based variance is negligible.

2.6 Variance of weighted estimators

The variance of weighted estimators may similarly be evaluated with respect to combinations of p, q or ξ, but not, of course, with respect to I. The two-phase approach is discussed in detail by SÄRNDAL and SWENSSON (1987) and LUNDSTRØM and SÄRNDAL (1999). There has also been discussion of conditional variances for weighted estimators. In particular, HOLT and SMITH (1979) make a case for conditioning the variance on the sample sizes within post-strata for a post-stratified estimator.

Chapter 3

Measuring the inflation of variances by non-response

3.1 neff: a general measure of variance inflation

An intuitive measure of the variance inflation due to non-response is the ratio between (a) the variance of the estimator based on the observed data (i.e. in the presence of non-response), and (b) the variance of the would-have-been estimator based on the complete data (i.e. in the hypothetical situation where non-response is absent). Clearly, estimators (a) and (b) should be specified in such a way that we may, as much as possible, attribute the difference between them to the fact that, while we observe all the values of interest in case (b), some of these values are concealed from us in case (a). Let θ be the population characteristic of interest. Let $\widehat{\theta}$ denote the observed estimator (a). Let $\tilde{\theta}$ be the conceptual estimator (b). The effect of non-response on variance, denoted by neff, is then

$$\text{neff} = \text{Var}(\widehat{\theta})/\text{Var}(\tilde{\theta}) \tag{3.1}$$

By definition the variance of $\tilde{\theta}$ is due to the sampling variation alone. In cases where neff > 1, the difference $\text{Var}(\widehat{\theta}) - \text{Var}(\tilde{\theta})$ can be considered as the additional component of variance causes by non-response. Moreover, the decomposition of $\text{Var}(\widehat{\theta})$ into two components due to, respectively, sampling and non-response, can sometimes be made explicit as we shall see later. However, artificial examples where $\text{Var}(\widehat{\theta}) < \text{Var}(\tilde{\theta})$ can easily be constructed, in which case $\text{Var}(\widehat{\theta}) - \text{Var}(\tilde{\theta})$ is negative and, thus, not a component of variance. We therefore define the neff in terms of a ratio.

It should also be noticed that definition (3.1) does not reflect uncertainty arising from lack of knowledge of the true non-response mechanism. The effect of non-response is to be evaluated under an assumed non-response model. While there exist Bayesian methods of inference which include extra model uncertainty (e.g. Rubin, 1987, and Forster and Smith, 1998), plausible interpretations, or analogies, from a frequentist perspective remain lacking.

The pair of estimators $(\widehat{\theta}, \tilde{\theta})$ is specified as follows. Let $s = \{1, \ldots, n\}$ denote the sample. Let y_i denote the vector of values of the survey variable associated with unit i. Let R_i be the non-response indicator for unit i, which takes the value 1 if the corresponding component of y_i is observed and 0 if it is missing. Let x_i be the vector of auxiliary values associated with unit i. We assume that x_i is not subject to non-response. Let $a_i = \pi_i^{-1}$ be the sampling weight, where π_i denotes the inclusion probability of unit i.

We shall consider both weighting adjustment and imputation based approaches of estimation. Consider first estimation based on imputation. Let y_s be the matrix with y_i as its ith column. Let $y_s^* = \{y_i^*; i \in s\}$ be the imputed matrix, where an element of y_s^* is identical to the corresponding element of y_s if it is observed. Let x_s be the matrix of auxiliary values similarly defined. Let $a_s = (a_1, \ldots, a_n)$ be the vector of sampling weights. We assume that the conceptual estimator $\tilde{\theta}$ is a function of a_s, y_s and x_s, denoted by

$$\tilde{\theta} = g(a_s, y_s, x_s). \tag{3.2}$$

The comparable imputation estimator $\widehat{\theta}$ is then defined as

$$\widehat{\theta} = g(a_s, y_s^*, x_s), \tag{3.3}$$

i.e. through the same function $g(.)$, but based on the imputed y_s^* instead of the conceptual y_s.

Next, consider weighting based on adjustment cells for unit non-response. Let R_s be the vector of non-response indicators in the sample. Let w_i be the adjusted weight if $R_i = 1$, which is given as

$$w_i = a_i \phi_i \ ,$$

where ϕ_i denotes the non-response adjustment weight either at the sample or population level. The observed estimator is a function of w_s, R_s, x_s and the observed columns of y_s denoted by

$$\widehat{\theta} = g(w_s, R_s, y_s, x_s) \ . \tag{3.4}$$

We assume that the function $g(.)$ is specified in a way which allows for the fact that some columns of y_s are missing, and the weight w_i is defined for respondents only. Typically, this can be achieved by operating with R_i, y_i and $R_i w_i$ instead of y_i and w_i, and make the convention that $R_i y_i = 0$ and $R_i w_i = 0$ if $R_i = 0$. The comparable conceptual estimator $\tilde{\theta}$ is then given by

$$\tilde{\theta} = g(w_s, R_s, y_s, x_s) \text{ where } \phi_s = R_s = 1 \ , \tag{3.5}$$

i.e. through the same function $g(.)$, and the non-response adjustment weight $\phi_i \equiv 1$ for all $i \in s$.

Example Direct weighting estimation.

Let θ be the population mean of y_i. For (3.4) and (3.5) we use

$$g(w_s, R_s, y_s, x_s) = \left(\sum_{i \in s} R_i a_i \phi_i y_i \right) \Big/ \left(\sum_{i \in s} R_i a_i \phi_i \right) \ .$$

Example Calibration estimation with respect to the known population totals of x_i, denoted by the vector X_U.

Let θ be the population total of some univariate y_i. For (3.4) and (3.5) we use

$$g(w_s, R_s, y_s, x_s) = X_U^T \widehat{\beta} + \sum_{i \in s} R_i a_i \phi_i (y_i - x_i^T \widehat{\beta}) \ ,$$

$$\text{where } \widehat{\beta} = \left(\sum_{i \in s} R_i a_i \phi_i x_i x_i^T \right)^{-1} \left(\sum_{i \in s} R_i a_i \phi_i x_i y_i \right) .$$

3.2 Variance inflation for two frameworks of inference

3.2.1 Design-based

Under the pure design-based framework, R_i is treated as a constant associated with unit i, just like y_i and x_i. The only stochastic variation comes from sampling. The general definition (3.1) is then

$$\text{neff} = \mathrm{V}_p(\widehat{\theta}) / \mathrm{V}_p(\tilde{\theta}), \tag{3.6}$$

where V_p denotes the variance with respect to the sampling distribution, denoted by p_s. We assume $n < N$ to avoid the trivial case of $\mathrm{V}_p(\widehat{\theta}) = \mathrm{V}_p(\tilde{\theta}) = 0$.

Example Direct weighting estimation under simple random sampling without replacement (*srs wor*).

Let θ be the population mean, so that $\tilde{\theta} = \overline{y}$, i.e. the sample mean, and $\widehat{\theta} = \overline{y}_1$, which is the observed sample mean. Let n_1 be the size of the response sample. We have

$$\mathrm{V}_p(\widehat{\theta}) = \mathrm{V}_p(\mathrm{E}_p[\,\widehat{\theta}\,|\,n_1\,]) + \mathrm{E}_p(\mathrm{V}_p[\,\widehat{\theta}\,|\,n_1\,]) = \mathrm{V}_p(\theta_1) + \sigma_1^2 \mathrm{E}_p(n_1^{-1} - N_1^{-1}) = \sigma_1^2 \mathrm{E}_p(n_1^{-1} - N_1^{-1}),$$

where θ_1 and σ_1^2 are the population mean and variance among those with $R_i = 1$, and $N_1 = \sum_{i \in U} R_i$ and $U = \{1, \ldots, N\}$ denotes the population. Let $\gamma = N_1/N$. We have

$$\text{neff} \approx \frac{AV_p(\widehat{\theta})}{\mathrm{V}_p(\tilde{\theta})} = \frac{\sigma_1^2[(n\gamma)^{-1} - (N\gamma)^{-1}]}{\sigma^2(n^{-1} - N^{-1})} = \frac{\sigma_1^2}{\sigma^2 \gamma}$$

where $AV_p(\widehat{\theta}) = \sigma_1^2[(n\gamma)^{-1} - (N\gamma)^{-1}]$ denotes approximate variance, and σ^2 is the population variance. Notice that so far we have not made any assumptions about the population vector $R_U = (R_1, \ldots, R_N)$, which however are necessary in order to assess the *neff*. In particular, under the assumption that $\sigma_1^2 = \sigma^2$, we may estimate γ by $\widehat{\gamma} = n_1/n$, to obtain

$$\text{neff} \approx \frac{1}{\widehat{\gamma}} = \frac{n}{n_1} > 1$$

which, quite reasonably, is the ratio between the complete sample size and the observed sample size.

Example Estimation following hot-deck imputation under *srs wor*.

Again, let θ be the population mean, so that $\tilde{\theta} = \overline{y}$ and $\widehat{\theta} = (n_1\overline{y}_1 + n_0\overline{y}_0^*)/n$ which is the imputed sample mean where $n_0 = n - n_1$ and $\overline{y}_0^*$ denotes the mean of the imputed values. Let s_1^2 be the sample variance among the respondents. It follows that

$$\mathrm{V}_p(\widehat{\theta}) = \mathrm{E}_p(\mathrm{V}_I[\,\widehat{\theta}\,|\,s\,]) + \mathrm{V}_p(\mathrm{E}_I[\,\widehat{\theta}\,|\,s\,]) = \mathrm{E}_p(n_0 s_1^2/n^2) + \mathrm{V}_p(\overline{y}_1)\ ,$$

where E_I and V_I denote expectation and variance with respect to the hot-deck imputation. We have

$$\text{neff} \approx \frac{AV_p(\widehat{\theta})}{\mathrm{V}_p(\tilde{\theta})} = \sigma_1^2 \left(\frac{1-\gamma}{n} + \frac{1}{n\gamma}\right) \Big/ \left(\frac{\sigma^2}{n}\right)$$

provided $n/N \approx 0$. Again, under the assumption that $\sigma_1^2 = \sigma^2$, we obtain with $\gamma = n_1/n$,

$$\text{neff} \approx 1 - \widehat{\gamma} + \frac{1}{\widehat{\gamma}} = 1 - \frac{n_1}{n} + \frac{n}{n_1} > 1$$

Notice that the non-response effect is larger than in the case of direct weighting estimator, because the hot-deck imputation constitutes an additional source of variation.

3.2.2 Quasi design-based

The quasi-design-based framework of inference is based upon a known sampling distribution and an assumed response distribution. Either data generation process may be considered to precede the other (see Section 2.4). The values of the survey variable, however, are still treated as constants.

Two-phase approach

In this case, non-response occurs conditional on the selected sample. The general definition (3.1) becomes

$$\text{neff} = \mathrm{V}_{pq}(\widehat{\theta})/\mathrm{V}_{pq}(\tilde{\theta}) = \mathrm{V}_{pq}(\widehat{\theta})/\mathrm{V}_p(\tilde{\theta}),$$

where V_{pq} denotes variance with respect to the sampling distribution and the assumed response distribution, and $\mathrm{V}_{pq}(\tilde{\theta}) = \mathrm{V}_p(\tilde{\theta})$ by definition. Provided $\widehat{\theta}$ is such that

$$\mathrm{E}_q(\widehat{\theta}|s) = \tilde{\theta} \quad , \tag{3.7}$$

the variance of $\widehat{\theta}$ admits an explicit decomposition for large samples. That is,

$$\mathrm{V}_{pq}(\widehat{\theta}) = \mathrm{E}_p(\mathrm{V}_q[\,\widehat{\theta}\,|\,s\,]) + \mathrm{V}_p(\mathrm{E}_q[\,\widehat{\theta}\,|\,s\,]) = \mathrm{V}_q(\,\widehat{\theta}\,|\,s\,) + \mathrm{V}_p(\tilde{\theta}),$$

where the first component in the approximate expression contains variations arising from non-response and the second component contains variation due to sampling alone. It follows that

$$\text{neff} \approx 1 + \frac{\mathrm{V}_q(\,\widehat{\theta}\,|\,s\,)}{\mathrm{V}_p(\tilde{\theta})} \tag{3.8}$$

Notice that the explicit decomposition does not apply exactly but approximately

for large samples. Moreover, the *neff* is necessarily larger or equal to unity under the current framework of inference.

The condition (3.7) is satisfied for several commonly used estimation methods under their respective assumptions of non-response. Including e.g. estimators based on known /observable adjustment cells, and estimators based on hot-deck or regression imputation. Indeed, exactly or approximately, it seems a reasonable requirement for any estimation method under the assumed non-response model. On the other hand, simulation studies (Särndal and Swensson, 1987, and Zhang, 2002), indicate that evaluation of the non-response effect (3.8) is fairly robust towards non-response model misspecification, so that the condition (3.7) may not be very critical for (3.8) in practice.

Example Direct weighting estimation under *srs wor* with observations missing completely at random (MCAR).

Let θ be the population mean, so that $\tilde{\theta} = \overline{y}$ and $\widehat{\theta} = \overline{y}_1$ as before. We have

$$\text{neff} \approx 1 + \frac{\mathrm{V}_q(\overline{y}_1|s)}{\mathrm{V}_p(\overline{y})} = 1 + \frac{\mathrm{E}_q(\mathrm{V}_q[\overline{y}_1|s,n_1])}{\sigma^2(n^{-1}-N^{-1})} = 1 + \frac{s^2\mathrm{E}_q(n_1^{-1}-n^{-1})}{\sigma^2(n^{-1}-N^{-1})},$$

where s^2 is the sample variance among y_s. Under the MCAR assumption, we obtain

$$\widehat{\text{neff}} = 1 + \frac{n_1^{-1}-n^{-1}}{n^{-1}-N^{-1}}$$

which reduces to n/n_1, i.e. $\widehat{\text{neff}}$ under the pure randomisation framework, provided $N^{-1} \approx 0$.

Example Estimation following hot-deck imputation under *srs wor* and MCAR mechanism.

Again, let θ be the population mean, so that $\tilde{\theta} = \overline{y}$, and $\widehat{\theta} = \overline{y}^* = (n_1\overline{y}_1 + n_0\overline{y}_0^*)/n$. The variance of $\tilde{\theta}$ remains the same, whereas

$$\begin{aligned}
\mathrm{V}_q(\widehat{\theta}|s) &= \mathrm{V}_q\left\{\mathrm{E}_q\left(\mathrm{E}_I\left[\overline{y}^*|s,n_1,y_1\right]|s,n_1\right)\right\} + \mathrm{E}_q\left\{\mathrm{V}_q\left(\mathrm{E}_I\left[\overline{y}^*|s,n_1,y_1\right]|s,n_1\right)\right\} \\
&\quad + \mathrm{E}_q\left\{\mathrm{E}_q\left(\mathrm{V}_I\left[\overline{y}^*|s,n_1,y_1\right]|s,n_1\right)\right\} \\
&= \mathrm{V}_q\left\{\mathrm{E}_q\left(\overline{y}_1|s,n_1\right)\right\} + \mathrm{E}_q\left\{\mathrm{V}_q\left(\overline{y}_1|s,n_1\right)\right\} + \mathrm{E}_q\left\{\mathrm{E}_q\left(\frac{n-n_1}{n^2}s_1^2\middle|s,n_1\right)\right\} \\
&= 0 + s^2\mathrm{E}_q\left(1/n_1 - 1/n\right) + s^2\mathrm{E}_q\left(1/n - n_1/n^2\right)
\end{aligned}$$

where the last term is due to the additional variation in hot-deck imputation.

Quasi design-based approach with population response mechanism

In this case, response precedes sampling. We have $\mathrm{V}_{q^U p}(\tilde{\theta}) = \mathrm{V}_p(\tilde{\theta})$ by definition, and

$$\mathrm{V}_{q^U p}(\widehat{\theta}) = \mathrm{E}_{q^U}(\mathrm{V}_p[\,\widehat{\theta}\,|\,R_U\,]) + \mathrm{V}_{q^U}(\mathrm{E}_p[\,\widehat{\theta}\,|\,R_U\,]) \approx \mathrm{V}_p(\,\widehat{\theta}\,|\,R_U\,) + \mathrm{V}_{q^U}(\mathrm{E}_p[\,\widehat{\theta}\,|\,R_U\,]) \quad (3.9)$$

Typically, however,

$$\mathrm{V}_{q^U}(\mathrm{E}_p[\widehat{\theta}|R_U])/\mathrm{V}_p(\widehat{\theta}|R_U) = O(n/N),$$

such that the *neff* is approximately the same as that under the pure randomization framework provided $n/N \approx 0$, but not otherwise.

Example Direct weighting estimation under *srs wor* and MCAR mechanism.

Let θ be the population mean, so that $\tilde{\theta} = \overline{y}$ and $\widehat{\theta} = \overline{y}_1$. We have

$$\mathrm{V}_{q^U}(\mathrm{E}_p[\overline{y}_1|R_U]) = \mathrm{V}_{q^U}(\widehat{\theta}_1|R_U) = \mathrm{E}_{q^U}(\mathrm{V}_{q^U}[\theta_1|N_1]|R_U) + \mathrm{V}_{q^U}(\mathrm{E}_{q^U}[\theta_1|N_1]|R_U) \approx \frac{\sigma^2}{N}(\gamma^{-1} - 1),$$

whereas $AV_p(\widehat{\theta}|R_U) = \sigma_1^2(n^{-1} - N^{-1})\gamma^{-1}$. It follows that $\widehat{\mathrm{neff}} = n/n_1$ if $n/N \approx 0$.

Example Estimation following hot-deck imputation under *srs wor* and MCAR mechanism.

Again, let θ be the population mean, so that $\tilde{\theta} = \overline{y}$, and $\widehat{\theta} = \overline{y}^* = (n_1\overline{y}_1 + n_0\overline{y}_0^*)/n$. The approximate variance $AV_p(\,\widehat{\theta}\,|\,R_U\,)$ is the same as previously given for the hot-deck imputation estimator, whereas

$$\mathrm{V}_{q^U}(\mathrm{E}_p[\,\widehat{\theta}\,|\,R_U\,]) = \mathrm{V}_{q^U}(\mathrm{E}_p\{\mathrm{E}_I[\,\widehat{\theta}\,|\,s, R_s\,]\,|\,R_U\,\}) = \mathrm{V}_{q^U}(\mathrm{E}_p[\,\overline{y}_1\,|\,R_U\,]) = \mathrm{V}_{q^U}(\theta_1).$$

Chapter 4

Variance estimation

4.1 Introduction

In the previous two chapters we have discussed the variances of point estimators in the presence of non-response. We now consider the estimation of these variances using observed data from responding units. We shall focus on the case when missing data have been replaced by imputed values and the point estimator is based upon these imputed data. Much of the discussion, especially in Section 4.2., will also be relevant to weighted point estimators.

We begin by recalling the alternative definitions of variances in Chapter 2, reflecting the alternative frameworks for inference. We now outline approaches to the estimation of these alternative variances and refer to some of the literature where methods have been developed. We employ the same notation as in Chapter 2 where the parameter is denoted θ and the estimator of θ based upon imputed data is denoted $\widehat{\theta}^*$.

Design-based $\widehat{\mathrm{V}}_p$ (or $\widehat{\mathrm{V}}_{pI}$)

Considering first deterministic imputation, we suppose that $\widehat{\theta}^*$ may be expressed as in (2.2). We may then apply any available variance estimation method for a smooth function of estimated totals, e.g. linearization or replication methods. The variance estimator may thus be constructed as a natural extension of the variance estimator which would be chosen in the absence of missing data. The main complexity involved is in obtaining the expression of form (2.2). See Shao and Steel (1999) for examples.

In the case of stochastic imputation, the variance is given in (2.3) and the same variance estimation approach may be applied to estimate the first term in (2.3) by obtaining an expression of form (2.2) for $\mathrm{E}_I(\widehat{\theta}^*)$. It is then necessary to add an estimator of $\mathrm{V}_I(\widehat{\theta}^*)$ (Shao and Steel, 1999, Section 5).

Quasi Design-based with Population Response Mechanism: $\mathrm{V}_{q^U p}$ (or $\mathrm{V}_{q^U pI}$)

The approach described in the previous section may be naturally extended, following the approach of Shao and Steel (1999), to estimate the variance in (2.4) or (2.5). The first term in (2.4) and (2.5) may be estimated in exactly the same way as the design-based variance. In cases where the sampling fraction is negligibly small the second term

in (2.4) or (2.5) may be omitted as an approximation. Otherwise, SHAO and STEEL (1999) discuss how this term may be estimated using linearization and the assumption about q^U that the R_i are independent for different i and that $\Pr(R_i = 1)$ is constant within imputation cells, with imputation conducted independently in different cells. The smaller the sampling fraction the more robust this method may be expected to be under departures from this assumption about q^U.

Quasi Design-based – two-phase sampling: V_{pq} (or V_{pqI})

The standard approach to variance estimation in this case is to make assumptions about the response mechanism, q, and then to use standard methods of variance estimation for two-stage sampling from the survey sampling literature. A common assumption is that the R_i are independently distributed and that $\Pr(R_i = 1)$ is constant (uniform response), either across the population or within imputation cells. See RAO and SITTER (1995); RAO (1996), and SITTER and RAO (1997) for linearization approaches. Replication approaches have also been proposed within this framework and under the same assumptions. See RAO and SHAO (1992); RAO and SITTER (1995); SITTER and RAO (1997), and YUNG and RAO (2000) for jackknife methods and SHAO *et al.* (1998) and RAO and SHAO (1999) for balanced repeated replication methods.

Model-anticipated variance $\mathrm{E}_\xi \mathrm{V}_{pq}$ (or $\mathrm{E}_\xi \mathrm{V}_{pqI}$)

Variance estimation methods for this case are generally based upon the model ξ, which was explicit or implicit in generating the imputed values and is called the *imputation model*, together with the assumption that y_i is missing at random (MAR) given the covariates used in imputation (this is also called an *unconfounded* missing data mechanism). Such methods therefore make stronger assumptions than the two-phase approach, in the sense that they depend upon a model, but weaker assumptions in the sense that the missing data mechanism only needs to be MAR not uniform. DEVILLE and SÄRNDAL (1994) set out the approach for regression imputation. This approach is discussed further by RANCOURT *et al.* (1994); LEE *et al.* (2002) and LUNDSTRØM and SÄRNDAL (2002).

4.2 Some specific methods for estimating variance inflation

In Chapter 3 we introduced the concept of *neff*, to measure the inflation of variance arising from non-response. In the notation of that chapter, we again use θ to denote the parameter of interest. We use $\widehat{\theta}$ to denote the estimator based upon the observed data, which may be the same as $\widehat{\theta}^*$ if the estimator is based upon imputed data. We use $\tilde{\theta}$ to denote the estimator based upon the complete data, in the hypothetical situation where no non-response had arisen. The *neff* measure was defined in (3.1) as the ratio of the variances of these two estimators.

In simulation studies the variances and *neff*, provided they are too complicated to be given in closed forms, can be approximated in a straightforward manner by the Monte Carlo method using independent replicates of $\widehat{\theta}$ and $\tilde{\theta}$. In real applications, and based on the selected sample and the observed data, we need to estimate the variances of $\widehat{\theta}$ and $\tilde{\theta}$, where the variances are defined according to a given framework for inference. In this

section we describe some general resampling methods, whose detailed properties remain to be examined by simulations. The notation will follow that in Chapter 3.

4.2.1 Variance estimation based on plug-in sample

Plug-in variance estimation

Suppose that either a consistent or unbiased variance estimator is available for $\mathrm{V}_p(\tilde{\theta})$ given the complete data, denoted by $\tau(a_s, y_s, x_s)$. A plug-in variance estimator is then given by

$$\widehat{\mathrm{V}}_p(\tilde{\theta}) = \tau(\tilde{a}_s, \tilde{y}_s, \tilde{x}_s),$$

for some suitable *plug-in* sample data $\tilde{s} = (\tilde{a}_s, \tilde{y}_s, \tilde{x}_s)$. The plug-in sample data are such that

$$(\tilde{a}_i, \tilde{y}_i, \tilde{x}_i) = (a_i, y_i, x_i) \quad \text{if } r_i = 1,$$

i.e. if there are no missing values on unit i. For any unit from which data are missing, we still observe (a_i, x_i) and, possibly, components of y_i. Nevertheless, we do not require the plug-in values to be the same as the observed ones in such cases, as long as they are plausibly chosen for the purpose of variance estimation. Hence the term plug-in sample rather than imputed sample.

Notice that estimation of $\mathrm{V}(\tilde{\theta})$ may be possible without constructing the entire plug-in sample, especially when closed variance formula is available given the complete data, such as when $\tilde{\theta}$ is the Horvitz-Thompson or generalised regression estimator (Särndal *et al*., 1992). In such cases, we only need to plug in certain complete-data statistics estimated from the observed data.

The situation is similar for $\mathrm{V}_q(\widehat{\theta} \,|\, a_s, y_s, x_s)$. In the first place, closed formulae are often available when $\widehat{\theta}$ is derived under either the MCAR or MAR assumption, such as examples earlier on direct weighting and hot-deck imputation. Otherwise, suppose that given y_s, $\mathrm{V}_q(\widehat{\theta}|a_s, y_s, x_s)$ can be evaluated straightforwardly by Monte Carlo approximation based on independent bootstrap replicates of $\widehat{\theta}$ under the assumed non-response model. A plug-in estimator is then given by applying the same bootstrap variance estimator to some suitable plug-in data sample data $\tilde{s}$.

Construction of the plug-in sample

For the purpose of variance estimation, the plausibility of a plug-in sample rests primarily on whether it exhibits similar sample variation to that of the complete sample.

Both weighting adjustment and imputation methods often assume that the units can be divided into so-called response homogeneity groups (RHG). Under the RHG model, response is independent between units, homogenous within the same response group, but heterogeneous across the groups. Let s_h denote the response groups, for $h = 1, \ldots, H$. Let $F_{1,h}(a, y, x)$ be the empirical distribution function (EDF) based on $\{(a_i, y_i, x_i); i \in s_{1,h}\}$, where $s_{1,h}$ denotes the set of complete cases within group h, which is an unbiased estimator

of the EDF based on $\{(a_i, y_i, x_i); i \in s_h\}$. A plausible plug-in set of data for units in group h which are subject to non-response, denoted by $\tilde{s}_{0,h}$, is such that

$$F_{0,h}(a, y, x) = F_{1,h}(a, y, x).$$

For discrete (a, y, x), we simply put $m_h(a, y, x)$ cases with values (a, y, x) into $\tilde{s}_{0,h}$, where

$$m_h(u, y, x)/\tilde{m}_h = n_h(a, y, x)/n_h$$

and $n_h(a, y, x)$ is the number of complete cases with values (a, y, x), and $n_h = \sum_{a,y,x} n_h(a, y, x)$, and $\tilde{m}_h$ is either the observed or estimated size of $\tilde{s}_{0,h}$. When at least some of the components of (a, y, x) are continuous, we may approximate $F_{1,h}$ by the following procedure:

1. choose a suitable set of grid values (a_k, y_k, x_k) for $k = 1, \ldots, K$;
2. divide $s_{1,h}$ into $s_{1,h,1}, \ldots, s_{1,h,k}$, such that $\{(a_i, y_i, x_i); i \in s_{1,h,k}\}$ are centered around (a_k, y_k, x_k), and let n_{hk} be the size of $s_{1,h,k}$;
3. fill into $\tilde{s}_{0,h}$ a simple random sample of size m_{hk}, drawn with replacement from $s_{1,h,k}$, together with the associated values (a_i, y_i, x_i), for $m_{hk}/\tilde{m}_h = n_{hk}/n_h$ and $k = 1, \ldots, K$.

In cases where the response probability is unique for individual values of (a, y, x) the following two procedures can be considered. Firstly, we may stratify the response probabilities into a suitable number of classes, and then construct the plug-in sample as described above for the RHG model. Secondly, let $\widehat{y}_i$ denote the estimated expected values which are missing. Variations among the conceptual y_s can be approximated by adding 'noises' to these expected values in a suitable manner.

Previous simulation studies (ZHANG, 2002) indicate that variance estimation is fairly robust towards to the various approximations involved in the construction of the plug-in sample. It would be interesting to check whether this robustness holds in a wider range of situations.

4.2.2 Variance estimation using bootstrap or jackknife

Based on complete data, both the Taylor linearization method and resampling methods, such as the jackknife and bootstrap, have been justified under the general stratified multistage sample design (for an overview see SHAO, 1996) provided that $\tilde{\theta}$ is some smooth (non-linear) function of plug-in estimates of a number of population characteristics. Given non-response, the linearization approach, as well as some of the resampling approach such as the rescaling bootstrap (RAO and WU, 1988), easily becomes intractable. For some theoretical results on bootstrap and jackknife in the presence of non-response we refer to SHAO and SITTER (1996); RAO and SHAO (1992); YUNG and RAO (2000); CHEN and SHAO (2001). The additional non-response assumptions considered in these works are either MCAR or MAR. Yet, as long as we can envisage $\widehat{\theta}$ as a smooth function of the

observed data, we may expect the validity of the proposed bootstrap and jackknife methods to carry through also for estimators derived under nonignorable non-response models. The key to success lies in appropriate treatment of non-response in the resamples. We shall not go into the details here. Instead we discuss below some points which are either work taking notice of, or require further investigations in the future.

First of all, the works cited above are somewhat vague when it comes to the exact nature of the framework of inference. Are they targeted at $\mathrm{V}_p(\widehat{\theta})$ or $\mathrm{V}_{pq}(\widehat{\theta})$? Of course, there is no difference between the two frameworks in the case of sampling with replacement. Moreover, as long as the sampling fractions of the first-stage PSU's are negligible, as in Rao and Shao (1992); Yung and Rao (2000) and Chen and Shao (2001), we may ignore the term $\mathrm{V}_{q^U}(\mathrm{E}_p[\,\widehat{\theta}\,|\,R_U\,])$ in (3.9), such that $\mathrm{V}_p(\widehat{\theta}) \approx \mathrm{V}_{q^U p}(\widehat{\theta})$, and the conceptual difference between them matters little in reality. But what about the cases where the sampling fractions are appreciable? It is instructive then to consider the cited bootstrap and jackknife methods in the extreme case of a census subject to non-response, where $\mathrm{V}_p(\widehat{\theta}) = 0$ and

$$\mathrm{V}_{q^U p}(\widehat{\theta}) = \mathrm{V}_q[\widehat{\theta}(R_U, a_U, y_U, x_U)|y_U, x_U, a_U = 1].$$

The jackknife methods require that the first-stage PSU's are either selected with replacement, or may be so treated. Although the condition is clearly not satisfied in the case of $s = U$, the jackknife variance estimator may still be calculated. Let $\widehat{\theta}_{(hi)}$ be the (hi)th jackknife replicate of $\widehat{\theta}$, i.e. the estimator obtained using the jackknife weights on deletion of the ith PSU in stratum h. Since $\widehat{\theta}_{(hi)}$ differs depending on which PSU is being deleted, the jackknife variance estimator will not be zero. But neither is it an estimate of $\mathrm{V}_q(\widehat{\theta}|y_U, x_U)$. Rather, in this case it is an estimate of the variance, w.r.t. to the unknown 'super-population' distribution of (R_U, y_U, x_U), because it is approximately the same as the bootstrap variance from resampling with replacement the PSU's in the population with all the associated values of (R_{hi}, y_{hi}, x_{hi}).

Unlike the jackknife, the bootstrap approach is not restricted to estimators of smooth functionals. Neither does it require sampling with replacement. Shao and Sitter (1996) recommend two resampling procedures given non-response, i.e. the without-replacement bootstrap BWO, Sitter (1992a), and the mirror-match bootstrap MMB, Sitter (1992b). In the extreme case of census subject to non-response, the BWO draws stratified resamples of the population size. Since these are drawn without replacement, the bootstrap replicates of $\widehat{\theta}$ are all identical to the observed one. It follows that the BWO is aimed at $\mathrm{V}_p(\widehat{\theta})$ in this case. The MMB is undefined in this case. We consider instead *srs wor* with $n_h = N_h - 1$, where the stratum sample size equals to the stratum population size minus one. A special MMB in this case amounts to stratified bootstrap with replacement, with the stratum resample size given by $N_h\,(N_h - 1)$. The bootstrap replicates of the estimator for the stratum population mean have then variance $s_h^2/\,[N_h\,(N_h - 1)]$, which matches the theoretical variance estimate under the pure randomization framework.

It is clear from the discussions above that we need to modify the existing bootstrap and jackknife methods for estimation of $\mathrm{V}_{q^U p}(\widehat{\theta})$ in cases where the sampling fractions are

appreciable. We propose the following procedure based on the BWO (SITTER, 1992a, Section 3.1):

1. Create the pseudopopulation as in the BWO. However, instead of the observed sample data (a_s, R_s, y_s, x_s), use some suitable plug-in sample $\tilde{s}$ as described in Subsection 4.2.1 above.

2. Generate the population response vector, denoted by R_U^*, under the estimated non-response model.

3. Resample as in the BWO, but with the response indicators generated in Step 2 above.

4. Derive $\widehat{\theta}^*$ based on the resample in the same way as $\widehat{\theta}$ on the observed sample.

5. Repeat Step 2-4 to generate independent bootstrap replicates of $\widehat{\theta}$, and calculate the standard Monte Carlo approximation to $\mathrm{V}_{q^U p}(\widehat{\theta})$.

In this way, the modified BWO contains an extra randomization (Step 2) with respect to the estimated response distribution. Notice that the pseudopopulation values of (y_U, x_U), generated in Step 1, are held fixed for all the resamples, which is appropriate for $\mathrm{V}_{q^U p}(\widehat{\theta})$.

4.3 A general jackknife method for imputed data

Some jackknife methods have already been referred to in Subsection 4.2.2. These methods tend to make specific assumptions about the sampling scheme. For example, RAO and SHAO (1992) assumed that at the first stage, the units (or clusters) are selected with replacement (or equivalently the first stage sampling fractions are small) and with equal probabilities. However, without replacement sampling is common in practice. The jackknife method set out in this section can be applied for unequal probability sampling without replacement, where the sampling fractions may be large, as in workpackage 6.

The two-phase approach described in Section 2.4 will be used as the framework. The selection of a sample s of size n from a finite population of size N by a sampling design $p(s)$ is treated as the first phase. The selection of a set of respondents $r \subset s$ by a probability mechanism $q(r|s)$ is treated as the second phase. We suppose that the item y is observed only for units in r. Let R_i be the response indicator for unit i, so that $R_i = 1$ if unit i responds to item y and $R_i = 0$ otherwise. Let $q_i = q(R_i = 1 \,|\, s)$ be the response probability for unit $i \in s$. We assume that the units respond independently of one another. We further assume uniform response, that is $q_i = q$ for all $i \in s$. This assumption will be relaxed in Subsection 4.3.4, where we allow for different response probabilities in different imputation classes.

4.3.1 A jackknife method for general sampling designs but no non-response

CAMPBELL (1980) proposed a generalized jackknife variance estimator that allows for complex sampling. BERGER and SKINNER (2003) show that this jackknife estimator is design-consistent for parameters of interest that can be expressed as a function of means.

Consider the HÁJEK (1981) estimator,

$$\widehat{\mu} = \sum_{i \in s} w_i y_i \ ,$$

of a population mean $\mu = N^{-1} \sum_{i \in U} y_i$, where

$$w_i = \pi_i^{-1} / \widehat{N},$$

π_i is the first-phase inclusion probability of unit i and $\widehat{N} = \sum_{i \in s} \pi_i^{-1}$. The generalized jackknife estimator of the variance of $\widehat{\mu}$ is given by

$$\widehat{\mathrm{V}} = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \varepsilon_{(i)} \varepsilon_{(j)} \ , \tag{4.1}$$

where the π_{ij} are the joint inclusion probabilities for the (first phase) sampling design. The quantities ε_i are pseudo-values given by

$$\varepsilon_{(j)} = (1 - w_j)(\widehat{\mu} - \widehat{\mu}_{(j)}), \tag{4.2}$$

where $\widehat{\mu}_{(j)}$ is the estimator of μ which has the same form as $\widehat{\mu}$, but is based only on the data that remain after omitting the j-th unit; that is

$$\widehat{\mu}_{(j)} = \sum_{i \in s, i \neq j} w_{i(j)} y_i \ ,$$

where

$$w_{i(j)} = \pi_i / \widehat{N}_{(j)} \ ,$$

with $\widehat{N}_{(j)} = \widehat{N} - \pi_j^{-1}$, $(i \neq j)$.

The weights w_i are more suitable than $(N\pi_i)^{-1}$ for the jackknife as, whenever a unit is deleted, these weights reduce to the jackknife weights for simple random sampling without replacement (*srs wr*). Indeed, whenever a unit j is deleted, the weights $w_{i(j)}$ reduce to the usual jackknife weight $(n-1)^{-1}$ under *srs wr*. The term $(1 - w_i)$ is a correction for unequal probabilities.

4.3.2 Deterministic imputation

Let r and $m = s - r$ denote, respectively, the sets of respondents and non-respondents to item y. When $i \in r$, the value of y_i is known and does not need to be imputed. When $i \in m$, the value of y_i is missing and needs to be imputed. Let y_i^* denote the imputed

value for missing $y_i, i \in m$. Let $\tilde{y}$ be the imputed item, where $\tilde{y}_i = y_i$ if $i \in r$ and $\tilde{y}_i = y_i^*$ if $i \in m$. We assume that imputed values are identifiable; that is, the response indicators R_i are known. The imputed estimator of the population mean μ is

$$\widehat{\mu}_I = \sum_{i \in s} w_i \tilde{y}_i. \tag{4.3}$$

We consider two forms of deterministic imputation for which (4.3) is asymptotically unbiased.

Mean imputation

First, consider the case of mean imputation where $y_i^* = \widehat{\mu}_r$ with

$$\widehat{\mu}_r = \sum_{i \in r} w_{i;r} y_i \quad , \tag{4.4}$$

where

$$w_{i;r} = w_i \left(\sum_{j \in r} w_j \right)^{-1} .$$

Suppose that whenever a responding unit $j \in r$ is deleted, the imputed values y_i^* are adjusted by an amount

$$\alpha_{i(j)} = y_{i(j)}^* - y_i^* \ , \tag{4.5}$$

where $y_{i(j)}^*$ is the value one would impute for the non-responding unit i. Under mean imputation, $\alpha_{i(j)} = \alpha_j$, for all $i \in m$, where

$$\begin{aligned} \alpha_j &= \widehat{\mu}_{r(j)} - \widehat{\mu}_r \ , \\ \widehat{\mu}_{r(j)} &= \sum_{i \in r, i \neq j} w_{i;r(j)} y_i \ , \\ w_{i;r(j)} &= w_i \left(\sum_{l \in r, l \neq j} w_l \right)^{-1} , \end{aligned} \tag{4.6}$$

where $j \neq i \in r$. Thus, the adjusted imputed values are $y_i^* + \alpha_j$. For *srs wr*, (4.6) is the adjustment proposed by RAO and SHAO (1992). The imputed values are unchanged whenever a non-responding unit is deleted.

The adjusted jackknife variance estimator proposed is

$$\widehat{\mathrm{V}}_a = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \varepsilon_{(i)}^a \varepsilon_{(j)}^a + \widehat{\mathrm{V}}_r \quad , \tag{4.7}$$

where

$$\widehat{\mathrm{V}}_r = (1-q) \sum_{i \in s} \pi_i {\varepsilon_{(i)}^a}^2 \tag{4.8}$$

$$\varepsilon_{(i)}^a = (1 - \tilde{w}_j)(\widehat{\mu}_I - \widehat{\mu}_{I(j)}^a), \tag{4.9}$$

where $\widehat{\mu}^a_{I(j)}$ is computed with adjusted imputed values and

$$\begin{aligned} 1-\tilde{w}_j &= 1-w_{j;r} && \text{if } j \in r, \\ &= (1-w_j)(1-q\pi_j)^{-0.5} && \text{if } j \in m. \end{aligned} \tag{4.10}$$

Note that with equal probabilities and small sampling fraction $1-\tilde{w}_j \approx 1$ for $j \in s$. For mean imputation, one can use any value for $1-\tilde{w}_j$ when $j \in m$, as $\widehat{\mu}_I - \widehat{\mu}^a_{I(j)} = 0$ when $j \in m$. However, in Subsection 4.3.3 we will see that the values of $1-\tilde{w}_j$ when $j \in m$ guarantee asymptotic unbiasedness with stochastic imputation.

BERGER and RAO (2004) show that (4.7) is consistent with uniform non-response. In (4.8) q needs to be estimated, for example by $\widehat{q} = \sum_{i \in r} w_i$. Here, we treat q as fixed.

The first term in (4.7) reduces to the adjusted jackknife estimator of RAO and SHAO (1992) for *srs wr*. The second term $\widehat{\mathrm{V}}_r$ in (4.7) is a correction for large sampling fractions. Indeed, in this case, the first term of (4.7) is small and $\widehat{\mathrm{V}}_r$ has a larger contribution to the variance. $\widehat{\mathrm{V}}_r$ equals zero for full response. With a census, the first term equals zero, as $\pi_{ij} - \pi_i\pi_j = 0$ and the variance is only due to the non-response: $\widehat{\mathrm{V}}_a = \widehat{\mathrm{V}}_r$. The following lemma gives conditions for $\widehat{\mathrm{V}}_r$ to be negligible.

Lemma $\widehat{\mathrm{V}}_r \widehat{\mathrm{V}}_a^{-1} \to 0$ *if* $\lambda_{\min} \to 1$ *and* $\lambda_{\max} \to 1$ *, where* $\lambda_{\min}$ *and* $\lambda_{\max}$ *are the minimum and the maximum value of* $\lambda_i = (1-\pi_i)(1-\pi_i q)^{-1}$ *for* $i \in U$.

The proof of this lemma is given in BERGER and RAO (2004). The jackknife variance estimator proposed depends on π_{ij}, which is often unknown. However, if $p(s)$ is a single stage stratified sampling design, we can use the HÁJEK (1964) variance estimator. This estimator can be easily computed for a wide range of sampling designs (BERGER, 1998) with any standard package using weighted least squares residuals.

Ratio imputation

Now suppose that values x_i of an auxiliary variable are available for all the sampled units $i \in s$ and that ratio imputation is employed with $y^*_i = x_i \widehat{\mu}_r / \widehat{\mu}_{x;r}$, where

$$\widehat{\mu}_{x;r} = \sum_{i \in r} w_{i;r} x_i.$$

Suppose that whenever a responding unit $j \in r$ is deleted, the imputed values y^*_i are adjusted by the amount $\alpha_{i(j)}$ defined by (4.5) with

$$y^*_{i(j)} = x_i \frac{\widehat{\mu}_{r(j)}}{\widehat{\mu}_{x;r(j)}},$$

where

$$\widehat{\mu}_{x;r(j)} = \sum_{i \in r, i \neq j} w_{i;r(j)} x_i.$$

Thus, the adjusted imputed values are $y^*_i + \alpha_{i(j)}$. The imputed values are unchanged whenever a non-responding unit is deleted. Note that unlike mean imputation, with ratio imputation the adjustment depends on i. BERGER and RAO (2004) show that (4.7) is consistent under ratio imputation with uniform non-response.

4.3.3 Stochastic imputation

Deterministic imputation leads to serious under-estimation for measure of population distributions. For example, mean imputation can under-estimate the population variance of an item y. This bias can be reduced by using stochastic imputation, such as hot deck.

Consider

$$y_i^* = \widehat{\mu}_r + e_i \ , \tag{4.11}$$

where the e_i are independent random variables such that $\mathrm{E}_I(e_i) = 0$, where $\mathrm{E}_I(.)$ denotes expectation with respect to the stochastic imputation specified by the distribution of e_i given s and r. The e_i can be generated from a parametric distribution. Alternatively, the e_i can be the residual $y_i - \widehat{\mu}_r$ of a donor $i \in r$ selected with replacement with selection probabilities $w_{i;r}$. The second option is called weighted random hot deck. Hot deck is more appealing for categorical variables.

Suppose that the missing values are imputed using (4.11) and consider an imputed estimator given by (4.3). Suppose that whenever a responding unit $j \in r$ is deleted, the imputed values y_i^* are adjusted by an amount

$$\alpha_{i(j)} = \mathrm{E}_{I(j)}(y_i^*) - \mathrm{E}_I(y_i^*), \tag{4.12}$$

where $\mathrm{E}_{I(j)}(.)$ denotes expectation with respect to the distribution of e_i given s and r after omitting the j-th sample unit. Note that (4.12) reduces to (4.5) when the e_i are not random. The adjusted jackknife estimator is given by (4.7) with

$$\varepsilon_{(j)}^a = (1 - \tilde{w}_j)(\widehat{\mu}_I - \widehat{\mu}_{I(j)}^a), \tag{4.13}$$

where $\widehat{\mu}_{I(j)}^a$ is computed after adjusting imputed values by (4.12) and $\tilde{w}_j$ is defined by (4.10). Note that $\widehat{\mu}_I - \widehat{\mu}_{I(j)}^a \neq 0$ for $j \in m$ and the term $(1 - q\pi_j)^{-0.5}$ in (4.10) is necessary to guarantee asymptotic unbiasedness with large sampling fractions (see BERGER and RAO, 2004). BERGER and RAO (2004) show that the adjusted jackknife estimator (4.7) is asymptotically unbiased under stochastic imputation and uniform non-response if the adjusted pseudo-values are given by (4.13).

4.3.4 Imputation classes

The uniform response assumption is often unrealistic in practice. It can be relaxed by forming $\nu \geqslant 2$ imputation classes and then assume uniform response within imputation classes. Let s_υ and r_υ denote the sample and the set of respondents in the ν-th class. Suppose that imputation is performed independently within each imputation class so that $y_i^* = \widehat{\mu}_{r\upsilon} + e_{i\upsilon}$; where $\mathrm{E}_{I\upsilon}(e_{i\upsilon}) = 0$, where $\mathrm{E}_{I\upsilon}(.)$ denotes the expectation with respect to s_υ and r_υ. The mean for the respondent of the ν-th class is

$$\widehat{\mu}_{r\upsilon} = \sum_{i \in r_\upsilon} w_i y_i \left(\sum_{j \in r_\upsilon} w_j \right)^{-1}$$

Thus, the imputed estimator $\widehat{\mu}_I$ is

$$\widehat{\mu}_I = \sum_{\upsilon} \frac{\widehat{N}_\upsilon}{\widehat{N}} \widehat{\mu}_{I\upsilon} \ ,$$

where

$$\widehat{\mu}_{I\upsilon} = \sum_{i \in s_\upsilon} w_i \tilde{y}_i \left(\sum_{j \in s_\upsilon} w_j \right)^{-1} .$$

4.4 A non-Bayesian approach to multiple imputation

Multiple imputation is a method specifically designed to enable variance estimation in the presence of missing data (Rubin; Rässler, 1987; 2004). The basic idea is to create m multiple imputed values for each missing value. Given these values, it is possible to compute m values of the point estimator and of the 'naïve' variance estimator, both of which treat the imputed values as real. A variance estimator may then be constructed from these values according to Rubin's combination formula. For this estimator to be valid, the imputation method must display an appropriate level of variability across multiple imputations. In the terminology of multiple imputation, the imputation method is required to be "proper" (Rubin, 1987). Unfortunately, the methods used for imputing for non-response in national statistical institutes (NSI's) very seldom if ever satisfy the requirement of being "proper". However, the idea of creating multiple imputations to measure the imputation uncertainty and use it for variance estimation and for computing confidence intervals is still of interest. The problem is then that Rubin's combination formula is no longer valid with the usual non-proper imputations used by NSI's. The reason being that the variability in non-proper imputations is too little and the within imputation component must be given a larger weight in the variance estimate. The problem is then to determine what this weight should be to give valid statistical inference, and also for what kind of non-response mechanisms and estimation problems it is possible to determine a simple combination formula not dependent on unknown parameters. This section suggests an approach for studying this problem.

In Subsection 4.4.1 an approach for determining the combination of the imputed completed data sets is suggested, with two illustrations. Subsection 4.4.2 takes up the problem of using the same combination rule for all estimation problems with a given imputation method and data & response model.

4.4.1 An approach for determining an alternative combination formula for variance estimation

Let $s = (1, \ldots, n)$ denote the sample, with $Y^{(s)}$ the matrix with columns given by the realized sample values $y_1, \ldots, y_n$ of random variables $Y_1, \ldots, Y_n$ under the model ξ. The objective is to estimate some parameter θ. Now, let y_{obs} be the observed part of $Y^{(s)}$, with r being the subset of responding units in s, of size n_r,

$$y_{obs} = (y_i : i \in r).$$

Let $\widehat{\theta}$ be the estimator based on the full sample data $Y^{(s)}$, with $\mathrm{Var}(\widehat{\theta})$ estimated by $\widehat{\mathrm{V}}(\widehat{\theta})$. For $i \in s-r$ we impute by some method to give an imputed value y_i^* and let $\tilde{Y}^{(s)}$ denote the complete data $(y_{obs}, y_i^*, i \in s-r)$. Based on $\tilde{Y}^{(s)}$, we have

$$\widehat{\theta}^* = \widehat{\theta}(\tilde{Y}^{(s)})$$
$$\widehat{\mathrm{V}}^* = \widehat{\mathrm{V}}(\widehat{\theta}^*)$$

Multiple imputation of m repeated imputations leads to m completed data-sets with m estimates $\widehat{\theta}_i^*, i = 1, \ldots, m$ and related variance estimates $\widehat{\mathrm{V}}_i^*, i = 1, \ldots, m$. The combined estimate is given by

$$\overline{\theta}^* = \sum_{i=1}^{m} \widehat{\theta}_i^*/m.$$

The within-imputation variance is defined as

$$\overline{\mathrm{V}}^* = \sum_{i=1}^{m} \widehat{\mathrm{V}}_i^*/m$$

and the between-imputation component is

$$B^* = \frac{1}{m-1} \sum_{i=1}^{m} (\widehat{\theta}_i^* - \overline{\theta}^*)^2.$$

The total estimated variance of $\overline{\theta}^*$ is then proposed to be

$$W = \overline{\mathrm{V}}^* + \left(k + \frac{1}{m}\right) B^* \ ,$$

where k is to be determined such that

$$\mathrm{E}(W) = \mathrm{Var}(\overline{\theta}^*). \tag{4.14}$$

RUBIN (1987) has shown that $k = 1$ can be used with proper imputations, which essentially means drawing imputed values from a posterior distribution in a Bayesian framework.

In general, one has to determine the terms in (4.14). One way to try and do this is to use double expectation, conditioning on y_{obs}, that is,

$$\mathrm{E}(W) = \mathrm{E}\{\mathrm{E}(W|Y_{obs})\}$$

$$\mathrm{Var}(\overline{\theta}^*) = \mathrm{E}\{\mathrm{Var}(\overline{\theta}^*|Y_{obs})\} + \mathrm{Var}\{\mathrm{E}(\overline{\theta}^*|Y_{obs})\}$$

Typically,

$$\mathrm{E}(\overline{\mathrm{V}}^*) \approx \mathrm{Var}(\widehat{\theta}) \tag{4.15}$$

and

$$\mathrm{E}(B^*|y_{obs}) = \mathrm{Var}(\widehat{\theta}^*|y_{obs}).$$

Hence, approximately

$$\mathrm{E}(W) = \mathrm{Var}(\widehat{\theta}) + \left(\mathrm{E}(k) + \frac{1}{m}\right) \mathrm{EVar}(\widehat{\theta}^*|Y_{obs}) \tag{4.16}$$

Moreover,

$$\mathrm{Var}(\overline{\theta}^*|y_{obs}) = \mathrm{Var}(\widehat{\theta}^*|y_{obs})/m$$

and

$$\mathrm{E}(\overline{\theta}^*|y_{obs}) = \mathrm{E}(\widehat{\theta}^*|y_{obs}).$$

This implies that

$$\mathrm{Var}(\overline{\theta}^*) = \frac{1}{m}\,\mathrm{E}\{\mathrm{Var}(\widehat{\theta}^*|Y_{obs})\} + \mathrm{Var}\{\mathrm{E}(\widehat{\theta}^*|Y_{obs})\} \tag{4.17}$$

From (4.15) and (4.16), the equation (4.14) becomes

$$\mathrm{Var}(\widehat{\theta}) + \mathrm{E}(k)\mathrm{EVar}(\widehat{\theta}^*|Y_{obs}) = \mathrm{Var}\{\mathrm{E}(\widehat{\theta}^*|Y_{obs})\},$$

which gives the following general expression for $\mathrm{E}(k)$:

$$\mathrm{E}(k) = \frac{\mathrm{VarE}(\widehat{\theta}^*|Y_{obs}) - \mathrm{Var}(\widehat{\theta})}{\mathrm{EVar}(\widehat{\theta}^*|Y_{obs})}. \tag{4.18}$$

For this to be of interest, k must be, at least approximately, determined independent of unknown parameters. In addition, one needs to check that (4.15) holds.

To illustrate how (4.18) can be used we shall consider two special cases with random non-response.

Example Estimating a population average with hot-deck imputation

Consider a simple random sample from a finite population of size N, where the aim is to estimate the population average $\overline{Y}$ of some variable y. We shall assume completely random non-response, MCAR in the terminology of Subsection 1.2.1. We note that MCAR means that the response indicators $R_1, \ldots, R_N$ are independent with the same response probability $q_{=}\Pr(R_i = 1)$. The imputation method is the hot-deck method, where y_i^* is drawn at random from y_{obs}, and the estimate is the sample mean. Let $\overline{y}_r$ be the observed sample mean and $\widehat{\sigma}_r^2 = (n_r - 1)^{-1} \sum_{i \in r} (y_i - \overline{y}_r)^2$ the observed sample variance. Then $\overline{Y}^*$ is the imputation-based sample mean for the completed sample, and the combined estimator is given by

$$\overline{\overline{Y}}^* = \sum_{i=1}^{m} \overline{Y}_i^*/m.$$

Let $\overline{Y}_s$ denote the sample mean based on a full sample. Then,

$$\mathrm{Var}(\overline{Y}_s) = \sigma^2 \left(\frac{1}{n} - \frac{1}{N}\right),$$

with $\sigma^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \overline{Y})^2$ being the population variance.

$$\mathrm{E}(\overline{Y}^*|y_{obs}) = \overline{y}_r \text{ and } \mathrm{Var}(\overline{Y}^*|y_{obs}) = \frac{n-n_r}{n^2} \cdot \frac{n_r-1}{n_r} \widehat{\sigma}_r^2$$

using the facts that

$$\mathrm{E}(Y_i^*|y_{obs}) = \overline{y}_r$$

and

$$\mathrm{Var}(Y_i^*|y_{obs}) = \frac{n_r-1}{n_r} \widehat{\sigma}_r^2.$$

Here,

$$\widehat{\mathrm{V}}^* = \widehat{\sigma}_*^2 \left(\frac{1}{n} - \frac{1}{N} \right),$$

where

$$\widehat{\sigma}_*^2 = (n-1)^{-1} \left[\sum\nolimits_r (y_i - \overline{y}^*)^2 + \sum\nolimits_{s-r} (y_i^* - \overline{y}^*)^2 \right].$$

It can be shown that

$$\mathrm{E}(\widehat{\sigma}_*^2|y_{obs}) = \sigma^2 \left(1 - \frac{1}{n_r}\right) \left(1 + \frac{n_r}{n(n-1)}\right) \approx \sigma^2,$$

and (4.15) holds.

We find, from (4.18),

$$\begin{aligned}
\mathrm{E}(k) &= \frac{\mathrm{Var}\left(\overline{Y}_r\right) - \sigma^2 \left(\frac{1}{n} - \frac{1}{N}\right)}{\mathrm{E}\left(\frac{n-n_r}{n^2} \cdot \frac{n_r-1}{n_r}\right) \mathrm{E}(\widehat{\sigma}_r^2 \,|\, n_r)} \\
&= \frac{\sigma^2 \left(\mathrm{E}\left(\frac{1}{n_r}\right) - \frac{1}{N}\right) - \sigma^2 \left(\frac{1}{n} - \frac{1}{N}\right)}{\mathrm{E}\left(\frac{n-n_r}{n^2} \cdot \frac{n_r-1}{n_r}\right) \sigma^2} \\
&\approx \frac{(1-q)/q}{1-q} = \frac{1}{q} \quad ,
\end{aligned}$$

which is satisfied approximately by letting

$$k = \frac{1}{1-f} \quad ,$$

where $f = (n-n_r)/n$ is the rate of non-response.

Example Estimating a regression coefficient with residual imputation We shall assume completely random non-response as in the previous example.

Model: $Y_i = \beta x_i + \varepsilon_i$, with $\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i$; $i = 1, \ldots, n$.

It is assumed that <u>all</u> x_i's are known, also in the non-response sample. The full data estimator of β is given by

$$\widehat{\beta} = \sum_{i=1}^{n} Y_i \Big/ \sum_{i=1}^{n} x_i.$$

The unbiased estimator of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{x_i} (y_i - \widehat{\beta} x_i)^2.$$

We shall consider residual regression imputation:

Let $\widehat{\beta}_r$ be the $\widehat{\beta}$ - estimate based on observed sample r. Define the standardized residuals

$$e_i = (y_i - \widehat{\beta}_r x_i)/\sqrt{x_i}, \text{ for } i \in r \ .$$

For $i \in s - r$, draw the value of e_i^* at random from the set of observed residuals $e_i, i \in r$, and the imputed y-value is given by

$$y_i^* = \widehat{\beta}_r x_i + e_i^* \sqrt{x_i}.$$

Let $X = \sum_{i \in s} x_i, X_r = \sum_{i \in r} x_i$ and $X_{nr} = \sum_{i \in s-r} x_i = X - X_r$. All considerations from now on are conditional on n_r and X_r, and we aim to determine k directly from (4.18). Define the proportion of the x - total in the non-response group to be:

$$f_X = X_{nr}/X.$$

We now have

$$\widehat{\beta}^* = \left(\sum\nolimits_r y_i + \sum\nolimits_{s-r} y_i^* \right) / X$$

$$\widehat{\sigma}_*^2 = (n-1)^{-1} \sum\nolimits_r x_i^{-1} (y_i - \widehat{\beta}^* x_i)^2 + \sum\nolimits_{s-r} x_i^{-1} (y_i^* - \widehat{\beta}^* x_i)^2.$$

In order to determine k from (4.14) we need to check the validity of (4.15) and derive the following quantities: $\mathrm{Var}(\widehat{\beta}^* | y_{obs}), \mathrm{E}(\widehat{\beta}^* | y_{obs})$ and $\mathrm{Var}(\widehat{\beta})$.

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2 / X.$$

Consider (4.15) which is equivalent to

$$\mathrm{E}(\widehat{\sigma}_*^2) \approx \sigma^2.$$

Let $\widehat{\beta}_{nr} = \sum\limits_{s-r} y_i^* / X_{nr}$, and $\widehat{\sigma}_{nr}^2 = (n_{nr} - 1)^{-1} \sum\limits_{s-r} x_i^{-1} (y_i^* - \widehat{\beta}_{nr} x_i)^2$. Here, $n_{nr} = n - n_r$. Then, after some algebra, one can express $\widehat{\sigma}_*^2$ in the following way:

$$\widehat{\sigma}_*^2 = \frac{1}{n-1} \left((n_r - 1)\, \widehat{\sigma}_r^2 + (n_{nr} - 1)\, \widehat{\sigma}_{nr}^2 + \frac{X_r X_{nr}}{X} (\widehat{\beta}_r - \widehat{\beta}_{nr})^2 \right)$$

In this case,

$$\mathrm{E}(Y_i^*|y_{obs}) = \widehat{\beta}_r x_i + \overline{e}\sqrt{x_i}, \text{ where } \overline{e} = \sum\nolimits_r e_i/n_r \ ,$$
$$\mathrm{Var}(Y_i^*|y_{obs}) = x_i s_e^2, \text{ where } s_e^2 = n_r^{-1} \sum\nolimits_r (e_i - \overline{e})^2.$$

Using this, it can be shown that

$$\mathrm{E}(\widehat{\sigma}_*^2) = \sigma^2 \left(1 - \frac{c_1}{n-1} - \frac{4c_2}{(n-1)n_r} - c_3 f \frac{n-1}{n \cdot n_r}\right)$$

where c_1, c_2, c_3 lie in the interval (0,1).

Hence, $\mathrm{E}(\widehat{\sigma}_*^2) \approx \sigma^2$ and (4.15) follows, at least for moderate and large n_r.

Next, we look at $\mathrm{Var}(\widehat{\beta}^*|y_{obs})$ and $\mathrm{E}(\widehat{\beta}^*|y_{obs})$:

We see that $\widehat{\beta}^* = (\widehat{\beta}_r X_r + \widehat{\beta}_{nr} X_{nr})/X$, and

$$\mathrm{E}(\widehat{\beta}_{nr}|y_{obs}) = \widehat{\beta}_r + \frac{\overline{e}}{X_{nr}} \sum_{s-s_r} \sqrt{x_i}$$
$$\mathrm{Var}(\widehat{\beta}_{nr}|y_{obs}) = s_e^2/X_{nr}.$$

This gives us

$$\mathrm{E}(\widehat{\beta}^*|y_{obs}) = \widehat{\beta}_r + \frac{\overline{e}}{X} \sum_{s-s_r} \sqrt{x_i}$$
$$\mathrm{Var}(\widehat{\beta}^*|y_{obs}) = \frac{X_{nr}}{X^2} s_e^2.$$

Next, we need to find $\mathrm{EVar}(\widehat{\beta}^*|y_{obs})$ and $\mathrm{VarE}(\widehat{\beta}^*|y_{obs})$. We have:

$$\mathrm{VarE}(\widehat{\beta}^*|y_{obs}) = \mathrm{Var}(\widehat{\beta}_r) + \frac{(\sum_{s-s_r} \sqrt{x_i})^2}{X^2} \mathrm{Var}(\overline{e}) + 2\frac{\sum_{s-s_r} \sqrt{x_i}}{X} \mathrm{Cov}(\widehat{\beta}_r, \overline{e}).$$

Using the Cauchy-Schwarz inequality,

$$\left(\sum_{i=1}^n a_i b_i\right)^2 \leqslant \sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2$$

with $a_i = \sqrt{x_i}$ and $b_i = 1$, we see that

$$\left(\sum_{s-r} \sqrt{x_i}\right)^2 \leqslant n_{nr} X_{nr}. \tag{4.19}$$

Now, after some algebra we find that $\mathrm{Cov}(\widehat{\beta}_r, \overline{e}) = 0$ and

$$\mathrm{Var}(\overline{e}) = \frac{\sigma^2}{n_r}\left(1 - \frac{\left(\sum_{s_r} \sqrt{x_i}\right)^2}{n_r X_r}\right) = (1 - d_1)\frac{\sigma^2}{n_r}, \quad 0 \leqslant d_1 \leqslant 1.$$

Moreover, from (4.19),

$$\frac{\left(\sum_{s-s_r}\sqrt{x_i}\right)^2}{X^2} = \frac{d_2 n_{nr} X_{nr}}{X^2}, \quad 0 \leqslant d_2 \leqslant 1.$$

Hence,

$$\mathrm{VarE}(\widehat{\beta}^*|y_{obs}) = \frac{\sigma^2}{X_r} + \frac{(1-d_1)d_2 n_{nr} X_{nr}}{X^2} \cdot \frac{\sigma^2}{n_r}$$

Next we find that

$$\mathrm{E}(s_e^2) = \sigma^2(1 - \frac{1}{n_r}) - \mathrm{Var}(\overline{e}) = \frac{\sigma^2}{n_r}(n_r + d_1 - 2)$$

which gives us

$$EVar(\widehat{\beta}^*|y_{obs}) = \frac{X_{nr}}{X^2} \cdot \frac{\sigma^2}{n_r}(n_r + d_1 - 2).$$

From (4.18),

$$\begin{aligned} k &= \frac{\frac{\sigma^2}{X_r} + \frac{\sigma^2}{n_r} \cdot \frac{(1-d_1)d_2 n_{nr} X_{nr}}{X^2} - \frac{\sigma^2}{X}}{\frac{\sigma^2}{n_r} \cdot \frac{X_{nr}}{X^2}(n_r + d_1 - 2)} \\ &= \frac{n_r X^2 - n_r X \cdot X_r + (1-d_1)d_2 n_{nr} X_{nr} X_r}{X_r X_{nr}(n_r + d_1 - 2)} \\ &\approx \frac{X}{X_r} + (1-d_1)d_2 \frac{n_{nr}}{n_r}. \end{aligned}$$

We note that if all $x_i = 1$, then $d_1 = d_2 = 1$. Now, with $f_X = X_{nr}/X$ being the proportion of the x-total in the non-response group and $f = n_{nr}/n$ the rate of non-response, we finally get, since typically $(1-d_1)d_2 \approx 0$,

$$k \approx \frac{1}{1-f_X} + (1-d_1)d_2 \frac{f}{1-f} \approx \frac{1}{1-f_X}$$

for usual x-values and non-response rates.

4.4.2 Question: can we use the same k for a given situation and imputation method, for all scientific estimands?

We try here to give a general approach to this problem. As an illustration we consider the first example in the previous section. For other situations and imputation methods, similar considerations should be studied.

In the example referred to, we found that for estimating the population mean with the sample mean,

$$k = \frac{1}{1-f}, \text{ with } f = n_{nr}/n, \text{ the non-response rate.} \tag{4.20}$$

The question is now: is this k valid for *other* estimation problems as well, using the same imputation method. The answer, in general, is NO. What is needed is to find conditions for (4.20) to be valid. From (4.18),

$$\mathrm{E}(k) = \frac{\mathrm{VarE}(\widehat{\theta}^*|Y_{obs}) - \mathrm{Var}(\widehat{\theta})}{\mathrm{EVar}(\widehat{\theta}^*|Y_{obs})} \quad .$$

In this case, the stochastic variables are (s, r), so an alternative notation is to use (s, r) instead of Y_{obs}. Hence, (4.18) becomes

$$\mathrm{E}(k) = \frac{\mathrm{VarE}(\widehat{\theta}^*|s,r) - \mathrm{Var}(\widehat{\theta})}{\mathrm{EVar}(\widehat{\theta}^*|s,r)} \tag{4.21}$$

One obvious requirement is that, at least approximately

$$\mathrm{E}(\widehat{\theta}^*|s) = \widehat{\theta}, \tag{4.22}$$

i.e. the imputed estimator should estimate the same parameter as $\widehat{\theta}$.

If we restrict attention to estimators that are linear in $(y_i : i \in s)$,

$$\widehat{\theta} = \sum_{i \in s} a_i(s) y_i \ , \tag{4.23}$$

then we have the following two results, which are proved in the Appendix.

Lemma Assume $\widehat{\theta}$ is given by (4.23). Then $\widehat{\theta}$ satisfies (4.22) if and only if $a_i(s) = a(s)$ for all $i \in s$.

i.e., $\widehat{\theta} = a(s) \sum_{i \in s} y_i = n a(s) \overline{y}_s$.

Theorem Assume $\widehat{\theta}$ is given by (4.23) and satisfies (4.22). Then $\mathrm{E}(k) = 1/q$ and $k = 1/(1-f)$.

Let us look at some special cases:

1. $a(s) = 1/n$, same as in the earlier example.
2. Regression coefficient for regression through the origin:

   $$\widehat{\beta} = \sum_{i=s} y_i / \sum_{i \in s} x_i.$$

   Here (4.22) is satisfied with $a(s) = 1/\sum_{i \in s} x_i$, and hence $k = 1/(1-f)$.

3. A case where (4.22) does not hold is regression coefficient in linear regression not through the origin:

$$\widehat{\beta} = \frac{\sum_{i \in s} (x_i - \overline{x}_s) y_i}{\sum_{i \in s} (x_i - \overline{x}_s)^2}.$$

Here, $a_i(s) = \dfrac{x_i - \overline{x}_s}{\sum_{j \in s} (x_j - \overline{x}_s)^2}$, which is not independent of i and one can show that $\mathrm{E}(\widehat{\beta}^* | s) \approx q\beta$ (exact $[(nq-1)/(n-1)]\beta$). Hence, for regular regression problems hot-deck imputation cannot work.

Obviously, when y is correlated with known x in non-response group, one should utilize this in the imputation regardless of the estimation problem one considers.

4.5 Variance estimation in the presence of weighting for non-response

The representation of sampling and non-response as two phases of sampling and the definition of the variance of a weighted estimator with respect to these phases was referred to in Section 2.6. Methods of variance estimation for two-phase sampling can therefore be applied to this case, under suitable assumptions about the non-response mechanism (Särndal and Swensson, 1987). The application of these ideas to a broad class of calibration estimators is discussed by Lundstrøm and Särndal (1999), and Lundstrøm and Särndal (2002).

The explicit two-phase approach requires the estimation of two components of variance, for sampling and for non-response, and this can be complicated with complex designs and estimators.

In some circumstances, it may be reasonable to employ a much simpler approach, which effectively ignores the non-response by treating the respondents as the sample obtained from the given sampling design. The idea is to employ a variance estimator which is valid for the given sampling design and weighted estimator in the absence of non-response. One example where this may be reasonable is for a stratified multistage design with small sampling fractions within strata. In this case, a common variance estimator is based upon the 'ultimate clusters' formed from sampled elements within primary sampling units (PSUs) within strata. The validity of this estimator depends upon approximating the without replacement sampling of the PSUs by with replacement sampling. This variance estimator will still be valid if the sample is also subject to unit non-response, provided non-response operates independently and in a common way between PSUs. Another example where this approach may be reasonable consists of a jackknife variance estimator based upon the deletion of PSUs.

Some care may be needed in judging whether standard variance estimators remain valid in the case of non-response. For example, the variance estimators proposed for calibration

weighting by DEVILLE and SÄRNDAL (1992) assume that the design-weighted estimates of the calibration totals are unbiased. This assumption may be unreasonable in the case of non-response, when the 'true' response probabilities are unknown. It may still be possible, however, to use appropriate 'single phase' variance estimators without estimating the two components required by the two-phase approach. Some discussion in the case of raking is provided in DACSEIS workpackage 6.

When population-level auxiliary information is used in weighting, it may be reasonable to consider estimating a conditional variance. The most well-studied case consists of post-stratification, where HOLT and SMITH (1979) argue that the variance should be conditional on the sample sizes within strata. ZHANG (2002) discusses the use of such conditional variance estimation for a particular weighting approach designed to handle non-ignorable non-response.

Appendix A

Proofs of results in Section 4.4.2

In order to prove the two results we need some facts:

(a) n_r is binomial (n, p_r)

(b) s_r given s, n_r is a simple random sample from s of size n_r

(c) $P(R_i = 1|s, n_r) = n_r/n$ and $P(R_i = 1, R_j = 1|s, n_r) = \dfrac{n_r}{n} \cdot \dfrac{n_r - 1}{n - 1}$ (follows from (b))

(d) $\mathrm{E}(Y_i^*|s, s_r) = \overline{y}_r$ ($\Rightarrow \mathrm{E}(Y_i^*|s, n_r) = \overline{y}_s \Rightarrow \mathrm{E}(Y_i^*|s) = \overline{y}_s$)

(e) $\mathrm{Var}(Y_i^*|s, s_r) = \dfrac{n_r - 1}{n_r}\, \widehat{\sigma}_r^2$

($\Rightarrow \mathrm{Var}(Y_i^*|s, n_r) = \dfrac{n_r - 1}{n_r}\, \widehat{\sigma}_s^2$, where $\widehat{\sigma}_s^2 = \dfrac{1}{n_s - 1} \sum_{i \in s} (y_i - \overline{y}_s)^2$ and $\mathrm{Var}(Y_i^*|s) \approx \widehat{\sigma}_s^2$)

Proof of Lemma

We get

$$\mathrm{E}(\widehat{\theta}^*|s) = \mathrm{E}\left(\sum_{i \in s_r} a_i(s) y_i + \sum_{i \in s - s_r} a_i Y_i^* \Big| s \right)$$

$$= \mathrm{E}_{s_r|s} \mathrm{E}\left(\sum_{i \in s_r} a_i(s) y_i + \sum_{i \in s - s_r} a_i(s) Y_i^* \Big| s, s_r \right)$$

$$=^{(d)} \mathrm{E}\left(\sum_{i \in s_r} a_i(s) y_i \Big| s \right) + \mathrm{E}\left(\sum_{i \in s - s_r} a_i(s) \overline{y}_r \Big| s \right)$$

The first term is:

$$\mathrm{E}\left(\sum_{i \in s_r} a_i(s) y_i \Big| s \right) = \mathrm{EE}\left(\sum_{i \in s_r} a_i(s) y_i \Big| s, n_r \right)$$

$$= \mathrm{EE}\left(\sum_{i\in s} a_i(s) y_i R_i \Big| s, n_r\right) = \mathrm{E}\left(\sum_{i\in s} a_i(s) y_i P(R_i = 1|s, n_r)\right)$$

$$=^{(c)} \mathrm{E}\left(\sum_{i\in s} a_i(s) y_i \frac{n_r}{n} \Big| s\right) =^{(a)} p_r \widehat{\theta}.$$

The second term is:

$$\begin{aligned}
\mathrm{E}\left(\sum_{i\in s-s_r} a_i(s)\overline{y}_r \Big| s\right) &= \mathrm{EE}\left(\sum_{i\in s-s_r} a_i(s)\overline{y}_r \Big| s, n_r\right) \\
&= \mathrm{EE}\left(\frac{1}{n_r}\sum_{i\in s-s_r}\sum_{j\in s_r} a_i(s) y_j \Big| s, n_r\right) \\
&= \mathrm{EE}\left(\frac{1}{n_r}\sum_{i\in s}\sum_{j\in s} a_i(s) y_j (1-R_i) R_j \Big| s, n_r\right) \\
&= \mathrm{E}\left(\frac{1}{n_r}\sum_{i\in s}\sum_{\substack{j\in s \\ j\neq i}} a_i(s) y_j (\mathrm{E}(R_j|s, n_r) - \mathrm{E}(R_i R_j|s, n_r))\right) \\
&=^{(c)} \mathrm{E}\left(\frac{1}{n_r}\sum_{i\in s}\sum_{\substack{j\in s \\ j\neq i}} a_i(s) y_j \left(\frac{n_r}{n} - \frac{n_r}{n}\cdot\frac{n_r - 1}{n-1}\right)\right) \\
&=^{(a)} \frac{1-p_r}{n-1}\sum_{i\in s}\sum_{\substack{j\in s \\ j\neq i}} a_i(s) y_j \\
&= \frac{1-p_r}{n-1}\,(n\overline{a}(s) n\overline{y}_s - \widehat{\theta}), \text{ where } \overline{a}(s) = \sum_{i\in s} a_i(s)/n.
\end{aligned}$$

This implies that

$$\mathrm{E}(\widehat{\theta}^*|s) = p_r\widehat{\theta} + \frac{1-p_r}{n-1}\,(n^2\overline{a}(s)\overline{y}_s - \widehat{\theta})$$

and (4.22) is equivalent to

$$p_r\widehat{\theta} + \frac{1-p_r}{n-1}\,(n^2\overline{a}(s)\overline{y}_s - \widehat{\theta}) = \widehat{\theta}$$

$$\Leftrightarrow \widehat{\theta}\left(1 + \frac{1-p_r}{n-1} - p_r\right) = \frac{1-p_r}{n-1}\,n^2\overline{a}(s)\overline{y}_s$$

$$\Leftrightarrow \widehat{\theta}\,\frac{n(1-p_r)}{n-1} = \frac{1-p_r}{n-1}\,n^2\overline{a}(s)\overline{y}_s$$

$$\Leftrightarrow \widehat{\theta} = n\overline{a}(s)\overline{y}_s = \overline{a}(s)\sum_{i\in s} y_i \text{ and the result follows.} \quad \square$$

Proof of Theorem

From the lemma, $\widehat{\theta} = a(s)\sum_{i\in s} y_i = na(s)\overline{y}_s$ and $\widehat{\theta}^* = a(s)\left(\sum_{i\in s_r} y_i + \sum_{i\in s-s_r} a_i y_i^*\right)$.

$$\begin{aligned}
\mathrm{E}(\widehat{\theta}^*|s,s_r) &=^{(d)} a(s)(n_r\overline{y}_r + (n-n_r)\overline{y}_r) = na(s)\overline{y}_r \\
\mathrm{Var}(\widehat{\theta}^*|s,s_r) &=^{(e)} [a(s)]^2(n-n_r)\frac{n_r-1}{n_r}\,\widehat{\sigma}_r^2
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{VarE}(\widehat{\theta}^*|s,s_r) &= \mathrm{Var}(na(s)\overline{y}_r) = \mathrm{EVar}(na(s)\overline{y}_r|s) + \mathrm{VarE}(na(s)\overline{y}_r|s) \\
&= \mathrm{E}n^2[a(s)]^2\mathrm{Var}(\overline{y}_r|s) + \mathrm{Var}\{na(s)\mathrm{E}(\overline{y}_r|s)\} \\
&= \mathrm{E}n^2[a(s)]^2\{\mathrm{E}_{n_r|s}\mathrm{Var}(\overline{y}_r|s,n_r) + \mathrm{Var}_{n_r|s}\mathrm{E}(\overline{y}_r|s,n_r)\} + \mathrm{Var}\{na(s)\mathrm{E}_{n_r|s}\mathrm{E}(\overline{y}_r|s,n_r)\} \\
&=^{(b)} \mathrm{E}n^2[a(s)]^2\left\{\mathrm{E}_{n_r}\left(\widehat{\sigma}_s^2\left(\frac{1}{n_r}-\frac{1}{n}\right)\right) + \mathrm{Var}_{n_r}(\overline{y}_s)\right\} + \mathrm{Var}\{na(s)\mathrm{E}_{n_r}\overline{y}_s\} \\
&= \mathrm{E}n^2[a(s)]^2\left\{\left(\widehat{\sigma}_s^2\mathrm{E}\left(\frac{1}{n_r}\right)-\frac{1}{n}\right)+0\right\} + \mathrm{Var}\{na(s)\mathrm{E}_{n_r}\overline{y}_s\} \\
&= n^2\left(\mathrm{E}\left(\frac{1}{n_r}\right)-\frac{1}{n}\right)\mathrm{E}[a(s)]^2\widehat{\sigma}_s^2 + \mathrm{Var}\;\widehat{\theta}.
\end{aligned}$$

Next,

$$\begin{aligned}
\mathrm{EVar}(\widehat{\theta}^*|s,s_r) &=^{(f)} \mathrm{E}\left\{[a(s)]^2(n-n_r)\frac{n_r-1}{n_r}\,\widehat{\sigma}_r^2\right\} \\
&= \mathrm{EE}\left\{[a(s)]^2(n-n_r)\frac{n_r-1}{n_r}\,\widehat{\sigma}_r^2\Big|s,n_r\right\} \approx \mathrm{E}\{[a(s)]^2(n-n_r)\mathrm{E}(\widehat{\sigma}_r^2|s,n_r)\} \\
&=^{(b)} \mathrm{E}[a(s)]^2\widehat{\sigma}_s^2(n-n_r) = \mathrm{EE}([a(s)]^2\widehat{\sigma}_s^2(n-n_r)|s) \\
&= n(1-p_r)\mathrm{E}[a(s)]^2\widehat{\sigma}_s^2.
\end{aligned}$$

We find now, from 4.21

$$\begin{aligned}
\mathrm{E}(k) &= \frac{\mathrm{VarE}(\widehat{\theta}^*|s,s_r) - \mathrm{Var}(\widehat{\theta})}{\mathrm{EVar}(\widehat{\theta}^*|s,s_r)} \\
&= \frac{n^2\left(\mathrm{E}\left(\frac{1}{n_r}\right)-\frac{1}{n}\right)\mathrm{E}[a(s)]^2\widehat{\sigma}_s^2}{n(1-p_r)\mathrm{E}[a(s)]^2\widehat{\sigma}_s^2} = \frac{n(\mathrm{E}(1/n_r)-1)}{1-p_r} \approx \frac{(1/p_r)-1}{1-p_r} = \frac{1}{p_r}. \quad \square
\end{aligned}$$

References

Berger, Y. G. (1998): Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* **74**, 149–168.

Berger, Y. G. and Rao, J. N. K. (2004): Adjusted jackknife for imputation under unequal probability sampling. Manuscript.

Berger, Y. G. and Skinner, C. J. (2003): Adjusted jackknife variance estimator for unequal probability sampling without replacement. Manuscript.

Campbell, C. (1980): A different view of finite population estimation. In *Proceedings of the Section on Survey Research Methods*, pp. 319–324. Alexandria, Virginia: American Statistical Association.

Chen, J. and Shao, J. (2001): Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association* **96**, 260–269.

Cochran, W. G. (1977): *Sampling Techniques*. Third edition. New York: Wiley.

Deville, J. C. and Särndal, C. E. (1992): Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Deville, J. C. and Särndal, C. E. (1994): Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics* **10**, 381–394.

Fay, R. E. (1991): A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference*, pp. 381–440. U.S. Bureau of the Census.

Fay, R. E. (1996): Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490–498.

Forster, J. J. and Smith, P. W. F. (1998): Model-based inference for categorial survey data subject to non-ignorable non-response (with discussion). *Journal of the Royal Statistic Society* **60**, 57–70.

Groves, R. M., Dillman, D. A., Eltinge, J. L. and Little, R. (2002): *Survey Nonresponse*. New York: John Wiley & Sons.

Hájek, J. (1964): Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491–1523.

Hájek, J. (1981): *Sampling from a Finite Population*. New York, Marcel Dekker.

Holt, D. and Smith, T. M. F. (1979): Poststratification. *Journal of the Royal Statistical Society, A* **142**, 33–46.

Lee, H., Rancourt, E. and Särndal, C. E. (2002): Variance estimation from survey data under single imputation. In *Survey Nonresponse*, eds R. M. Groves, D. A. Dillman, J. L. Eltinge and R. Little. New York: John Wiley & Sons.

Little, R. J. A. and Rubin, D. B. (2002): *Statistical Analysis with Missing Data*. Second edition. New York: John Wiley & Sons.

Lundstrøm, S. and Särndal, C. E. (1999): Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics* **15**, 305–327.

Lundstrøm, S. and Särndal, C. E. (2002): *Estimation in the Presence of Nonresponse and Frame Imperfections*. Statistics Sweden.

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.

Rancourt, E., Särndal, C. E. and Lee, H. (1994): Estimation of the variance in the presence of nearest neighbour imputation. In *Preceedings of the Section on Survey Research Methods, American Statisitcal Association*, pp. 888–893.

Rao, J. N. K. (1996): On variance estimation with imputed survey data. *Journal of the American Statistical Association* **91**, 499–506.

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

Rao, J. N. K. and Shao, J. (1999): Modified balanced repeated replication for complex survey data. *Biometrika* **86**, 403–415.

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

Rao, J. N. K. and Wu, C. F. J. (1988): Resampling inference with complex survey data. *Journal of the American Statistical Association* **83**, 231–241.

Rässler, S. (2004): *The Impact of Multiple Imputation for DACSEIS*. DACSEIS Research Paper number 5.

Rubin, D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Särndal, C. E. and Swensson, B. (1987): A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* **55**, 279–294.

Särndal, C. E., Swensson, B. and Wretman, J. (1992): *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J. (1996): Resampling methods in sample surveys (with discussion). *Statistics* **27**, 203–254.

Shao, J., Chen, Y. and Chen, Y. (1998): Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association* **93**, 819–831.

Shao, J. and Sitter, R. R. (1996): Bootstrap for imputed survey data. *Journal of the American Statistical Association* **91**, 1278–1288.

Shao, J. and Steel, P. (1999): Variance estimation with composite imputation and non-negligible sampling fractions. *Journal of the American Statistical Association* **94**, 254–265.

Sitter, R. and Rao, J. N. K. (1997): Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics* **25**, 61–75.

Sitter, R. R. (1992a): A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755–765.

Sitter, R. R. (1992b): Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics* **20**, 135–154.

Skinner, C. J. and Rao, J. N. K. (2002): Jackknife estimation for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference* **102**, 149–167.

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

Zhang, L. C. (2002): A method of weighting adjustment for survey data subject to nonignorable nonresponse. Technical report, DACSEIS Research Paper Series No. 2.