Introduction to the design and analysis of complex survey data

Statistical Science
2017, Vol. 32, No. 2, 165–175
DOI: 10.1214/17-STS614

Introduction to the Design and Analysis of Complex Survey Data

Chris Skinner and Jon Wakefield

Abstract. We give a brief overview of common sampling designs used in a survey setting, and introduce the principal inferential paradigms under which data from complex surveys may be analyzed. In particular, we distinguish between design-based, model-based and model-assisted approaches. Simple examples highlight the key differences between the approaches. We discuss the interplay between inferential approaches and targets of inference and the important issue of variance estimation.

Key words and phrases: Design-based inference, model-assisted inference, model-based inference, weights, variance estimation.

1. INTRODUCTION

Sampling has proved an essential tool over the last century to enable society to collect wide ranging accurate information about populations through cost-efficient survey data collection. Moreover, sample surveys, together with experiments, have provided core methods of data collection to support the development and application of modern statistical methods to scientific research. The central roles of surveys and sampling are seeing some challenges in the twenty-first century and "big data" era (e.g., Japec et al., 2015). Big data is taken here to refer to data sources which are generated as secondary outcomes of existing systems (rather than as a result of designed primary data collection) and which cover 100% of units to which the system applies (and thus involve no sampling). Such sources can be cheap and can provide information much more rapidly. Also, since they cover 100% of units, they may provide more granular estimates. In addition to competition from such sources, sample surveys now face threats to their accuracy from increasing nonresponse and major cost pressures. Nevertheless, they continue to have essential roles that big data sources cannot replace. Survey variables can be designed for research questions of interest rather than these questions having to be adapted to the, often very limited, sets of variables available from big data sources. Samples can be designed to represent populations of interest rather than study populations having to be adapted to the typically selective coverage of big data sources. In the light of such essential roles, the sample survey continues to be the method of choice in many settings and this special issue seeks to reflect the continuing vitality of developments in statistical methods in this field. We also aim to capture some of the evolution of the field as it advances. In an ideal situation, survey data can provide an important complement to alternative data sources. For example, estimation methods which combine carefully collected survey data and "big" data, have the potential to leverage the advantages of both.

The design and analysis of surveys are fascinating enterprises. Unless one is trained in the field, however, they can be exercises shrouded in mystery. For instance, the expression, "a weighted analysis is recommended", is a standard accompaniment to public release survey datasets but, unfortunately, weights can be constructed in many different and subtle ways which can leave the uninitiated scratching their heads in bewilderment. In this special issue, we hope to provide some enlightenment, beginning in this opening paper with a gentle introduction to the central themes of complex survey analysis.

The complexity of survey data alluded to in the title of this paper refers to the complex nature of sampling

Chris Skinner is Professor, Department of Statistics, London School of Economics and Political Science, London, United Kingdom (e-mail: c.j.skinner@lse.ac.uk). Jon Wakefield is Professor, Departments of Statistics and Biostatistics, University of Washington, Seattle, Washington, USA (e-mail: jonno@uw.edu).

designs, involving, for example, stratification and multistage sampling, together with associated complications such as nonresponse. Although we shall provide a brief outline of complex *designs* at the end of this section, we focus in this paper on the methodology of estimation and analysis and the question of how to account for the complex sampling design, largely treated here as given.

A further source of mystery for the many secondary analysts of complex survey data is that, whereas many novel developments in the methodology of survey design and estimation have been introduced by the agencies conducting large surveys, not all the information relating to these developments is made available when data are released for general analysis. For example, survey weights and imputed values may be made available but some of the complex features of the sampling frame, design and "raw" data underlying such released information may remain concealed.

Historically, survey sampling has often been seen as a rather separate topic to much of the rest of statistics. Not only have survey design and survey data analysis typically been undertaken by different people but the estimation methodology associated with survey design has often been centered on the *design-based* (or *randomization*) approach, quite separate from the *model-based* inference featuring in much mainstream applied statistics. While the former approach has a clear rationale, it can be confusing to those who have received a conventional model-based training in statistics. The slow rate of inclusion of complex survey methods in much applied statistical software for analysis has also contributed to this separation. Nevertheless, we have recently sensed a greater degree of cross-fertilization of ideas between survey sampling and applied statistical methods of analysis. A key purpose of this paper and special issue is to help support the sharing of such ideas by opening up developments in the survey sampling literature to a broader readership.

Before proceeding to consider inference, however, we set the scene by outlining some features of complex designs. We only consider *probability sampling* in which the design is characterized via a probability distribution over the possible samples that may be collected. In particular, each unit in the population of interest has a nonzero probability of being selected. A complex design may be viewed as one deviating from the simplest design, *simple random sampling* (SRS), in which all subsets of n from N units are equally likely. Here, N denotes the size of the population which is sampled and n denotes the sample size. Why would one wish to deviate from SRS and, in particular, from one of its properties, that each unit in the population is selected with equal probability? One reason is for efficient estimation. Tillé and Wilhelm (2017) refers to the "false intuition that a sample must be similar to a population" and explains how more efficient estimation can often be achieved by sampling units with unequal probabilities. Other reasons include practical constraints imposed by the nature of frames from which the sample must be drawn and variable costs of data collection. See Valliant, Dever and Kreuter (2013) for a detailed account of a wide range of designs used in practice.

We now describe some basic designs. We emphasize that in practice these often act as the building blocks of more complex designs, because of the characteristics of the sampling frame or the population. A common design is *stratified simple random sampling* in which a group label is available for each unit in the population, and SRSs are taken within each of the groups. For example, in studies of individuals, the groups may correspond to demographic strata and geographical regions. Stratified random sampling can provide appreciable gains in efficiency if the variables defining the strata are associated with the response. The main impediment to its use is the availability of strata information on all members of the population. In *single-stage cluster sampling*, the population is again partitioned, but this time into what are called "clusters", or primary sampling units (PSUs). The PSUs are often defined geographically.

The key difference between cluster sampling and stratified sampling is that only a sample, rather than all, of the clusters are selected and then information is obtained from all individuals within clusters. Cluster sampling in general reduces efficiency, because of within-cluster correlation, but logistically it is very convenient, particularly in nationwide sampling. In *two-stage cluster sampling*, random samples are taken within the sampled clusters (PSUs). In large national surveys, *stratified multistage cluster sampling* (in which cluster sampling and stratified SRS is executed) is the norm, since it balances efficiency, logistical constraints, and the requirement for estimates of sufficient precision to be obtained for subgroups of interest.

For the remainder of this paper, we turn to inference, but see Tillé and Wilhelm (2017) for additional probability sampling designs, with an emphasis on new developments. In Section 2, we introduce and compare

design-based and model-based inference for complex survey data. Section 3 brings in the role of auxiliary information and describes *model-assisted* inference in which working models are adopted to suggest estimators/designs, but with the design-based approach being followed for inference. Sections 4 and 5 considers model parameters (as opposed to finite population characteristics) and nonprobability sampling. The important topic of variance estimation is the subject of Section 6. Final remarks conclude the paper in Section 7.

2. INFERENTIAL OVERVIEW

In this section, we provide a brief overview of inferential approaches to the analysis of survey data, assuming that probability sampling is used and there is no nonresponse. First, we define some notation. Let $y_k, k = 1, \ldots, N$ represent the values of a survey variable of interest on all N units in a well-defined finite population (e.g., all individuals aged 18 or greater in a particular administrative region); we shall also write $k \in U$ to index this collection. A sample of these units, denoted $S \subset U$, is taken via some probabilistic mechanism, where $p(s)$ denotes the probability of selecting $S = s$ with $\sum_s p(s) = 1$, and the values of y_k are only observed for $k \in S$. The probability of unit k being selected is $\pi_k = \sum_{s:k\in s} p(s)$, and the so-called *design-weight* is defined as $d_k = \pi_k^{-1}$. This weight is often interpreted loosely as the number of population units "represented" by the kth sampled unit.

We suppose here that finite population characteristics are the targets of inference. Other targets are considered in Section 4. It is common in the survey sampling literature to emphasize first the estimation of population totals, before moving on to targets which are functions of totals. There are a variety of reasons for this, beyond the obvious one that totals may indeed be targets of inference. One reason is that issues of bias can be dealt with more simply with totals than, say, means, especially when the population size N is unknown, as it often is. However, to present basic ideas in this paper we shall treat the finite population mean $\overline{y}_U = \frac{1}{N}\sum_{k\in U} y_k$ as our target of inference, since it is a more natural "unit-level" parameter of interest in much survey data analysis and includes, for example, a proportion as a special case.

We next contrast two broad paradigms under which inference based on survey data may be performed: the *design-based* and *model-based* approaches. These refer to two different sources of randomness, either from the randomization associated with probability sampling or from a model assumed to generate the population values y_k. Inference based on models is likely to be familiar to most readers and so we leave it till second. First, we discuss design-based inference, sometimes also called *randomization-based* analysis. Lohr (2010) is a popular text that is primarily concerned with design-based inference. The latter is more distinctive to survey sampling, though inference based on randomization is sometimes used in randomized experiments (Cox, 2006, Chapter 9).

2.1 Design-Based and Model-Based Approaches to Inference

2.1.1 *Design-based.* The population values $y_1, \ldots, y_N$, are viewed as fixed constants, with the collection of units selected, S, treated as random. Assuming N known, a standard weighted estimator of the population mean is

$$\overline{y}_{\mathrm{HT}} = \frac{\sum_{k\in S} d_k y_k}{N}. \tag{1}$$

This will be referred to as the HT estimator, since its numerator was proposed by Horvitz and Thompson (1952) for estimating the population total $\sum_{k\in U} y_k$. The primary motivation for such design weighting is to remove bias, as discussed in detail in Haziza and Beaumont (2017). Bias and other moments are evaluated in the design-based framework with respect to repeated sampling of units from the finite population U. We write $E_{\mathrm{S}}[\overline{y}_{\mathrm{HT}}]$, with the subscript S on the expectation emphasizing that we are averaging over possible subsets that could have been selected. Similarly, the variance of the estimator will be written as $\mathrm{var}_{\mathrm{S}}(\overline{y}_{\mathrm{HT}})$. We informally define two particular criteria: an estimator is *design unbiased* if its expectation (over all possible samples) is equal to the true value, and an estimator is *design consistent* if both the design bias and the variance go to zero as the sample size increases. For the latter, one must consider a sequence of populations, with the finite population size and the sample size tending to infinity.

An alternative estimator of the mean, defined whether N is known or not, is the Hájek estimator (Hájek, 1971):

$$\overline{y}_{\mathrm{HJ}} = \frac{\sum_{k\in S} d_k y_k}{\widehat{N}}, \tag{2}$$

where $\widehat{N} = \sum_{k\in S} d_k$, vindicating d_k's interpretation earlier as the number of population units represented by the kth sampled unit. The estimator (2) is biased

in finite samples, but is design consistent. The Hájek estimator is often preferred to the HT estimator even if N is known, and we give some rationale for this later when we consider models. This estimator illustrates another surprising aspect of traditional survey sampling, its preoccupation with the estimation of *ratios*. But many functions of interest (the mean, for example!) can be expressed as a ratio.

The key to deriving the properties of design-based estimators is to define binary indicators of selection I_k, such that $E_s[I_k] = \pi_k$, $\text{var}_s(I_k) = \pi_k(1-\pi_k)$, $E_s[I_k I_l] = \pi_{kl}$ and $\text{cov}_s(I_k, I_l) = \pi_{kl} - \pi_k\pi_l = \Delta_{kl}$, for $k, l \in U$. Here, the π_{kl} are the probabilities that units k and l are both selected; these are the key quantities required to determine the variances of estimators, as we demonstrate below. It is usual for designs to sample *without replacement* and for $\pi_{kl} \neq \pi_k\pi_l$ (this is in stark contrast to the model-based approach in which values are usually assumed to be independent, since they are drawn from a hypothetical *infinite* population). Desirable criteria from a design-based perspective are design unbiased (or design consistent) estimators with low variance.

It is straightforward to show that the design weighting in the HT estimator (1) does indeed remove bias:

$$E_s[\overline{y}_{HT}] = \frac{1}{N}E_s\Big[\sum_{k\in S} d_k y_k\Big] = \frac{1}{N}E_s\Big[\sum_{k\in U} d_k I_k y_k\Big]$$
$$= \frac{1}{N}\sum_{k\in U} \pi_k^{-1} E_s[I_k] y_k = \overline{y}_U.$$

The trick in the above derivation is to introduce the binary random variables I_k, and consequently sum over U; before that point the sum is over units in the random set S.

The unbiasedness arises because of the inverse probability weighting, a technique that is now in common use beyond survey sampling, particularly to adjust for nonresponse (Seaman and White, 2013). A key point is that we require $\pi_k > 0$ for the estimator to be design unbiased. This makes complete sense, because we cannot hope to achieve an unbiased estimator of a finite population characteristic, if some of the units can never be sampled.

The form of the variance of the HT estimator also follows straightforwardly:

$$\text{var}_s(\overline{y}_{HT}) = \frac{1}{N^2}\text{var}_s\Big(\sum_{k\in S} d_k y_k\Big)$$
$$= \frac{1}{N^2}\text{var}_s\Big(\sum_{k\in U} d_k I_k y_k\Big) \tag{3}$$
$$= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} d_k d_l \,\text{cov}_s(I_k, I_l) y_k y_l$$
$$= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

An unbiased estimator of the variance is

$$\widehat{\text{var}}_s(\overline{y}_{HT}) = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \tag{4}$$

Despite their ease of derivation, the forms of the variances in (3) and (4) can be quite mysterious to those raised in the model-based camp, since they do not appear to depend on the variances of the responses, y_k (but see the end of this subsection for the emergence of a familiar form). A pivotal requirement in the derivation of (4) is that $\pi_{kl} > 0$, that is, all pairs of units must have a positive probability of being selected. Although unbiased, the estimator in (4), as well as the closely related Sen–Yates–Grundy estimator given in Tillé and Wilhelm (2017), has some undesirable properties. For example, they can be negative. More importantly, in practice, the π_{kl} are often not available for all $k, l \in S$. This is usually the case for multistage, clustered designs, for example. In order to perform design-based inference, it is usual therefore to adopt alternative variance estimators. Approximations obtained by treating the design as "with replacement" are widely used, since the variance estimator is always nonnegative and it is not necessary to know π_{kl} for all $k, l \in S$ (see, e.g., Lohr, 2010). The use of resampling methods, such as the jackknife or bootstrap, is also common. Section 6 provides a fuller discussion.

To illustrate some of the expressions above, consider SRS for which

$$p(s) = \begin{cases} \binom{N}{n}^{-1}, & \text{if } s \text{ has } n \text{ elements,} \\ 0, & \text{otherwise,} \end{cases}$$
$$\pi_k = \frac{n}{N}, \quad d_k = \frac{N}{n}, \quad \pi_{kl} = \frac{n}{N}\frac{n-1}{N-1}.$$

We find $\widehat{N} = \sum_{k\in S} d_k = N$ so that $\overline{y}_{HT} = \overline{y}_{HJ} = \sum_{k\in s} y_k/n$ and the variance is $\text{var}_s(\overline{y}_{HT}) = (1-\frac{n}{N})\frac{S_y^2}{n}$ where $S_y^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \overline{y}_U)^2$ and $1-\frac{n}{N}$ is the *finite population correction* (if sampling from a hypothetical infinite population, this term would be 1). A design unbiased estimator of S_y^2 is

$$s_y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \overline{y}_{HT})^2, \tag{5}$$

so that we recover a familiar form for the variance of the estimator, albeit with a finite population correction.

2.1.2 *Model-based*. Under this, more mainstream, statistical approach the y_k are treated as realized values of random variables $Y_k, k = 1, \ldots, N$, which follow some specified model, viewing the population as drawn from a hypothetical infinite superpopulation. Frequentist evaluation refers now to repeated realizations from the model.

In conventional statistical modeling methods of data analysis, model parameters are typically of interest, rather than finite population characteristics, such as the finite population mean discussed above. Here, however, we consider how inference for the latter can be carried out by reference to the modeling framework. We now write the target of inference as $\overline{Y}_U = \frac{1}{N}\sum_{k\in U} Y_k$ to convey that it is random, but we emphasize that it represents the same target of inference as $\overline{y}_U$ in the design-based framework. Since the target is a random variable, the standard frequentist estimation rules do not apply and we refer to a *predictor*, rather than an *estimator* of $\overline{Y}_U$. The classic reference is Valliant, Dorfman and Royall (2000). Chambers and Clark (2012) provide an introduction. There are two main criteria considered to evaluate a predictor, denoted $\widehat{\overline{Y}}$. First, is the bias, $E_{\mathrm{M}}[\widehat{\overline{Y}} - \overline{Y}_U]$, where now both $\widehat{\overline{Y}}$ and $\overline{Y}_U$ are random. Second, the variance of the predictor with respect to the model is $E_{\mathrm{M}}[(\widehat{\overline{Y}} - \overline{Y}_U)^2]$. These criteria are the same as those used when random effects are predicted in a frequentist framework. The model-based approach can also be formulated in a Bayesian framework with inference about $\overline{Y}_U$ based on its posterior distribution given the data. We do not have the space to discuss this here, but the interested reader can consult, for example, Gelman (2007) and Little (2013).

As a simple illustration of the prediction approach, consider a model for which:

$$\mu = E_{\mathrm{M}}(Y_k), \quad \mathrm{var}_{\mathrm{M}}(Y_k) = \sigma^2, \tag{6}$$

with Y_k and Y_l independent. The predictor $\widehat{\overline{Y}}$ is taken as the sample mean $\widehat{\overline{Y}}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$. Our change of notation to $k = 1, \ldots, n$ acknowledges that the set of units S selected from N is no longer relevant. The sample mean is an unbiased predictor since

$$E_{\mathrm{M}}[\widehat{\overline{Y}}_n - \overline{Y}_U] = \frac{1}{n}\sum_{k=1}^{n} E_{\mathrm{M}}[Y_k] - \frac{1}{N}\sum_{k=1}^{N} E_{\mathrm{M}}[Y_k] = 0.$$

The prediction variance is

$$E_{\mathrm{M}}[(\widehat{\overline{Y}}_n - \overline{Y}_U)^2] = E_{\mathrm{M}}\left[\left(\frac{1}{n}\sum_{k=1}^{n} Y_k - \frac{1}{N}\sum_{k=1}^{N} Y_k\right)^2\right] = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}.$$

Substitution of σ^2 by the finite population variance, S_y^2, gives the same variance as obtained earlier for design-based inference under SRS.

It may appear that the complex sampling scheme plays no role in the model-based approach. There are, in fact, two fundamental ways in which it does. First, a complex sampling scheme will depend on the structure of the population through, for example, stratification or clustering. It is essential that this structure is captured in the model, for example using fixed effects for strata and random effects for clusters, if model-based inference is to be valid.

Second, in conventional model-based inference as above, it is assumed that the model specified at the population level also applies to all sample observations, however, they are sampled. This makes a strong implicit modelling assumption. Prediction under the model-based approach conditions on the selection indicators I_k. The assumption that the population model applies to the sample is therefore equivalent to assuming that the conditional distribution of Y_k given $I_k = 1$ is the same as its conditional distribution given $I_k = 0$. Sampling is then said to be *noninformative*. If this assumption does not hold, that is if $E_{\mathrm{M}}[Y_k|I_k = 1] \neq E_{\mathrm{M}}[Y_k|I_k = 0]$, where the subscript M indicates that the expectations are under a model, there is the potential for bias to arise, so-called *selection bias*. A key advantage of probability sampling is that it may be used to ensure the independence of I_k and Y_k, and hence to protect against selection bias. It is a risky endeavor to carry out inference from a nonprobability sample without such protection.

2.2 Switching the Paradigms

The two approaches may be compared and contrasted by examining design-based estimators using model-based criteria and vice-versa. Consider first the model bias of the HT and Hájek estimators under a model, as in (6), where $\mu = E_{\mathrm{M}}[Y_k]$ and the y_k in (1) and (2) are replaced by Y_k. Under this model,

$$E_{\mathrm{M}}[\overline{y}_{\mathrm{HT}} - \overline{Y}_U] = \left(\sum_{k=1}^{n} d_k - N\right)\frac{\mu}{N},$$

$$E_{\mathrm{M}}[\overline{y}_{\mathrm{HJ}} - \overline{Y}_U] = 0,$$

so that the Hájek estimator is always unbiased under this model but the HT estimator is only unbiased under the condition $\widehat{N} = \sum_{k=1}^{n} d_k = N$. This condition often does hold, for example, under SRS when $d_k = N/n$, but there may be problems in the performance of the HT estimator for sampling designs where it does not. An example is a design in which the sample size n is random and the weights are constant (i.e., $d_k = d$).

We switch now to consider the design bias of the model-unbiased estimator $\widehat{\overline{Y}}_n$, where Y_k is replaced by y_k. We have

$$E_{\mathrm{S}}[\widehat{\overline{Y}}_n] = E_{\mathrm{S}}\left[\frac{1}{n}\sum_{k=1}^{N} I_k y_k\right] = \frac{1}{n}\sum_{k=1}^{N} \pi_k y_k$$

and so this estimator will generally be design biased. Haziza and Beaumont (2017), Section 3, show that the design bias will only disappear if the π_k are uncorrelated with the y_k, which corresponds to the notion of noninformative sampling discussed in Section 2.1.2. This illustrates the (design) bias impact of model misspecification. The unweighted estimator $\widehat{\overline{Y}}_n$ is unbiased for $\overline{Y}_U$ under the model $\mu = E_{\mathrm{M}}[Y_k]$ (and the assumption of noninformative sampling) but does not protect against bias if these assumptions fail, unlike the design weighted (HT and Hájek) estimators.

So far as the variance is concerned, if we assume that $\mathrm{var}_{\mathrm{M}}(Y_k) = \sigma_k^2$ and that Y_k and Y_l are independent under the model, then the variance under the model of, for example, the Hájek estimator may be expressed as

$$E_{\mathrm{M}}[(\overline{y}_{\mathrm{HJ}} - \overline{Y}_U)^2] = \sum_{k=1}^{n}(\widehat{N}^{-1}d_k - N^{-1})^2\sigma_k^2 + \sum_{k=n+1}^{N} N^{-2}\sigma_k^2$$

(the same expression holds for the HT estimator with $\widehat{N}$ replaced by N). This expression will almost certainly not equal (3).

Studying the model variance of design-based estimators may help in assessing their efficiency. Consider, for example, the model

$$Y_k = \theta\pi_k + \varepsilon_k, \tag{7}$$

where the error terms ε_k are independent with $E_{\mathrm{M}}[\varepsilon_k] = 0$ and $\mathrm{var}_{\mathrm{M}}(\varepsilon_k) = \pi_k^2\sigma^2$, $k = 1, \ldots, n$. The weighted least squares estimator $\widehat{\theta}$, that minimizes

$$\sum_{k=1}^{n} \frac{(y_k - \theta\pi_k)^2}{\pi_k^2\sigma^2},$$

corresponds to $\overline{y}_{\mathrm{HT}}$, the HT estimator. Model (7) is not required to be correct for the properties of the HT estimator to be valid, but it does suggest situations in which we would expect the estimator to perform well (or not). This could be viewed as a *model-assisted* approach, which we discuss in more detail in Section 3. The widely cited Basu elephant example (Basu, 1971, Hájek, 1971) provides an extreme example in which the HT estimator performs poorly in a situation in which the responses y_k are not related to the sampling probabilities π_k. Briefly, a fictional circus owner would like to estimate the weight of his 50 strong herd of elephants, based on measuring a single elephant. Drawing on 3-year old records, he proposes to measure the weight of Sambo, an elephant who was previously of average weight, and multiply this weight by 50. This purposive design traumatizes the circus statistician, who is obsessed with using a design-unbiased estimator. To this end, he proposes an alternative plan in which Sambo is selected with probability $\pi_k = \frac{99}{100}$, and one of the remaining 49 creatures with probability $\pi_k = \frac{1}{100} \times \frac{1}{49}$. Letting y denote the weight of the selected elephant, the HT estimator is dy where $d = \frac{100}{99}$ if Sambo is selected, and $d = 100 \times 49$ if any of the other elephants is selected. Clearly, whichever elephant is selected, this estimator is unsatisfactory. Putting aside the wisdom of an $n = 1$ design, is it clear here that, by construction, y_k is not proportional to π_k. For further discussion, see Brewer (2002) and Lumley (2010), page 149.

From a modeling perspective, the Hájek estimator (2) arises from the model with $E_{\mathrm{M}}[Y_k] = \theta$ and $\mathrm{var}_{\mathrm{M}}(Y_k) = \pi_k^2\sigma^2$, $k = 1, \ldots, n$, and so one would expect it to outperform the HT estimator when the response is approximately constant (as opposed to being proportional to the sampling probabilities), which argues for its use in many instances, regardless of whether N is known,

To summarize, it is informative to view model-based estimators from a design-based perspective and vice versa, since it gives insight into situations in which the respective estimators will perform well. Conclusions we have drawn here include that the HT estimator may be design-unbiased, but biased with respect to particular models, and the model variance may not correspond to the design variance.

3. AUXILIARY VARIABLES AND MODEL-ASSISTED ESTIMATION

In most survey settings, auxiliary information about the population units will be available to assist both de-

sign and inference. From a predictive model-based perspective, it is very natural and commonplace to include auxiliary variables as covariates in regression models. This enables more efficient predictions to be achieved. Conditioning on auxiliary variables used in the design also helps to ensure that sampling is noninformative, that is, that I_k is independent of Y_k conditional on these covariates; hence it helps avoid the kinds of selection bias mentioned in Section 2.1.2.

From a design-based perspective, it is also possible to include auxiliary variables to improve inference, in particular, inference may be "assisted" by consideration of suitable covariates in a regression model for Y_k. The definitive text on model-assisted inference in surveys is Särndal, Swensson and Wretman (1992). For simplicity, suppose we know the mean $\overline{x}_U$ for a single variable x_k that we believe is associated with y_k. Consider the "working model",

$$Y_k = B_0 + B_1 x_k + \varepsilon_k, \tag{8}$$

where the error terms ε_k are independent with $E_{\mathrm{M}}[\varepsilon_k] = 0$, $\mathrm{var}_{\mathrm{M}}(\varepsilon_k) = \sigma^2$, $k = 1, \ldots, N$ and the intercept and slope are defined with respect to the finite population:

$$\begin{aligned} B_1 &= \frac{\sum_{k=1}^N (x_k - \overline{x}_U)(y_k - \overline{y}_U)}{\sum_{k=1}^N (x_k - \overline{x}_U)^2} \\ &= \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \frac{R S_y^2}{S_x^2}, \end{aligned} \tag{9}$$

$$B_0 = \overline{y}_U - B_1 \overline{x}_U, \tag{10}$$

where R, S_x and S_y are the correlation, standard deviation of x and standard deviation of y, in the population.

This model motivates $\widehat{B}_0 + \widehat{B}_1 \overline{x}_U$ as an estimator of $\overline{y}_U$, where $\widehat{B}_0$ and $\widehat{B}_1$ are design-based estimators of B_0 and B_1. For simplicity, consider SRS and the estimators:

$$\begin{aligned} \widehat{B}_1 &= \frac{\sum_{k \in S} (x_k - \overline{x}_S)(y_k - \overline{y}_S)}{\sum_{k \in S} (x_k - \overline{x}_S)^2} = \frac{r s_y^2}{s_x^2}, \\ \widehat{B}_0 &= \overline{y}_S - \widehat{B}_1 \overline{x}_S, \end{aligned}$$

where r is the sample correlation, and $\overline{y}_S, \overline{x}_S$ and s_x, s_y are, respectively, the means and standard deviations of x and y in the sample. Since ratios are involved, these estimators are not design unbiased for B_0 and B_1 but they are design consistent. The resulting estimator of $\overline{y}_U$ is

$$\widehat{B}_0 + \widehat{B}_1 \overline{x}_U = \overline{y}_S + \widehat{B}_1 (\overline{x}_U - \overline{x}_S).$$

This is the traditional *regression estimator* and is design consistent for $\overline{y}_U$ under SRS since $\overline{y}_S$ and $\overline{x}_S$ are design consistent for $\overline{y}_U$ and $\overline{x}_U$, respectively. The rationale for using the regression estimator rather than the simple estimator $\overline{y}_S$ is that it improves precision. It does this because, under (8) with $B_1 \neq 0$, the error $\overline{y}_S - \overline{y}_U$ is correlated with $\overline{x}_S - \overline{x}_U$. Under SRS, the large-sample variance of the regression estimator is given by

$$\left(1 - \frac{n}{N}\right) \frac{S_e^2}{n},$$

where $S_e^2 = \frac{1}{N-1} \sum_{k \in U} e_k^2$ with $e_k = y_k - (B_0 + B_1 x_k)$ denoting the residual. By comparison with the expression before (5) we see that the use of auxiliary information has reduced the variance by a factor of approximately R^2, the squared correlation between the y_k and x_k.

For general complex designs, we may use design weights in the estimators $\widehat{B}_0$ and $\widehat{B}_1$ and obtain, as discussed by Breidt and Opsomer (2017), the *generalized regression* (GREG) model assisted estimator of $\overline{y}_U$ as

$$\overline{y}_{\mathrm{GREG}} = \widehat{B}_0 + \widehat{B}_1 \overline{x}_U = \overline{y}_{\mathrm{HT}} + \widehat{B}_1 (\overline{x}_U - \overline{x}_{\mathrm{HT}}),$$

so that the HT estimator of the mean of the y_k is adjusted via the difference between the population mean of the x_k and its sample estimator. The GREG estimator is design consistent with finite sample design bias, but for large samples its precision will be greater than that of the HT estimator. The estimators $\widehat{B}_0$ and $\widehat{B}_1$ can also include weighting for heteroskedasticity in model (8) giving, for example, when B_0 is taken as zero, the widely used ratio estimator, as a special case (Breidt and Opsomer, 2017).

The overall model-assisted approach has a similar flavor to robust estimation using sandwich variance estimators, where a working model is specified, but the consistency of the estimator is guaranteed, under very weak assumptions, and in particular consistency does not depend on strong modeling assumptions. Breidt and Opsomer (2017) provide a much fuller account of model-assisted inference, including a wide range of extensions of the GREG approach.

We note that GREG and related estimators can be represented as weighted estimators, where the weights extend the simple idea of design weights introduced earlier by incorporating auxiliary population information. Various adjustments can be made and the construction of weights can be complex; the relevant issues are discussed in this issue by Haziza and Beaumont (2017) and Chen et al. (2017).

4. MODEL PARAMETERS

The previous two sections have focussed on the estimation of the finite population mean. Extensions to other finite population targets of inference can often be achieved by treating them as explicit functions of means and by estimating these targets by plugging the estimators of the means into the function; see, for example, Breidt and Opsomer (2017), Section 1. In this section, we focus instead on inference about model parameters, which raises additional issues. For a more detailed discussion, see Lumley and Scott (2017).

From a model-based perspective, if the model for the survey variables incorporates both the parameters of interest, θ say, and the auxiliary variables used in the design, as required to ensure the sampling scheme is noninformative, then it may be possible to treat the sampling scheme as ignorable for inference about θ and to employ standard unweighted approaches to inference. In this case, not only the sampling design but also the finite population may effectively play no role and the only model requiring consideration is that assumed to generate the sample data. However, a problem with conditioning on the design variables is that it may move the model away from that which is of interest. See, for example, the discussion of Lumley (2010), page 105, in relation to regression. Skinner, Holt and Smith (1989) refer to the distinction between conditioning or not on the design variables as "disaggregated" versus "aggregated" analyses, and note that the two approaches may serve quite different analytic purposes.

From a design based perspective, one may begin with the even more fundamental question of how to define the parameter of interest. A common approach is again to specify a (superpopulation) model of interest, including a parameter θ say, but where this is only treated as a "working model" for motivation and where the model will not be used for inference. With this purpose in mind, a "census parameter", θ_U may be defined, which is some estimator of θ, were the whole finite population to be observed by conducting a hypothetical census. For example, suppose we are interested in modelling unemployment and that a parameter θ of this model represents the probability that a person in the labour force in a particular population of people U is unemployed at a particular point of time. Then we might define θ_U as the actual proportion of the labour force in this population who are unemployed at this time. Under suitable modelling assumptions, it may be expected that θ_U will be a close approximation to θ if the population size is large. Taking θ_U as the parameter of interest (rather than θ) is attractive since it is a finite population quantity and so one may make design-based inference about it directly (as we did in Section 3 in the context of linear regression). This kind of approach is discussed in Lumley and Scott (2017), particularly in the context of pseudolikelihood estimation (Binder, 1983) which provides the basis of most established statistical packages for survey analysis.

The definition of a census parameter in terms of a specific estimation approach is somewhat arbitrary, however (Skinner, 2003), and it is often still preferable to take a model parameter θ as the target. In this case, it may still be reasonable to take as a point estimator $\widehat{\theta}$ the same estimator that would be used for θ_U but it will be necessary to modify variance estimation and related inference procedures by combining design-based and model-based inference.

For example, suppose we wish to make inference about $\mu = E_{\mathrm{M}}[Y]$, the mean in the (infinite) superpopulation from which the population of size N was sampled. We may treat the finite population mean $\overline{y}_U$ as the census parameter and, starting from a design based perspective, take the HT estimator $\overline{y}_{\mathrm{HT}}$ as a point estimator of both $\overline{y}_U$ and μ. Not only is it design unbiased for $\overline{y}_U$ but it is also unbiased for μ with respect to a joint design/model-based framework. Thus,

$$E[\overline{y}_{\mathrm{HT}}] = E_{\mathrm{M}}\big[E_{\mathrm{S}}[\overline{y}_{\mathrm{HT}}]\big] = E_{\mathrm{M}}[\overline{Y}_U] = \mu$$

(where we have replaced $\overline{y}_U$ by $\overline{Y}_U$ to emphasize that it is being treated as random). Turning to the variance, we need to consider the uncertainty due not only to the selection of a sample of size n but also due to the selection of the population of size N from the superpopulation. For simplicity, suppose the design is SRS. We obtain

$$\begin{aligned}\operatorname{var}(\overline{y}_{\mathrm{HT}}) &= E_{\mathrm{M}}\big[\operatorname{var}_{\mathrm{S}}(\overline{y}_{\mathrm{HT}})\big] + \operatorname{var}_{\mathrm{M}}\big(E_{\mathrm{S}}[\overline{y}_{\mathrm{HT}}]\big)\\ &= E_{\mathrm{M}}\left[\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}\right] + \operatorname{var}_{\mathrm{M}}(\overline{Y}_U)\\ &= \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n} + \frac{\sigma^2}{N} = \frac{\sigma^2}{n}.\end{aligned}$$

In this case, the variance is just as if a random sample was drawn from the superpopulation. Moreover, if N is much larger than n then the second term may be negligible and it may be argued that design-based inference suffices in practice. For discussion of inference about model parameters in more general settings, see Graubard and Korn (2002).

5. NONPROBABILITY SAMPLING AND NONRESPONSE

So far, we have assumed probability sampling and no nonresponse. In practice, non-response arises in most surveys of human populations and response rates have seen a relentless decline in many countries in recent decades. It thus becomes essential that inferential methods allow for missing data from nonresponse. Alongside the decline in response rates have been significant changes in survey practice, such as greater use of nonprobability sampling (Elliott and Valliant, 2017) associated, particularly, with web surveys; see Schonlau and Couper (2017).

Nonresponse can be with respect to items and/or the whole unit, and we consider only the latter. Nonresponse and nonprobability sampling share a common challenge for inference. They both involve forms of sample selection which are not fully under the control of the survey designer and to proceed, both require modeling assumptions. Elliott and Valliant (2017) provide an in-depth discussion of two broad approaches in this context, and we here briefly introduce these ideas.

One approach is *quasi-randomisation*, which seeks to represent the sample as if it had been obtained from probability sampling. In the case of nonresponse, this often involves treating the nonresponse as a second phase of sampling, as discussed by Haziza and Beaumont (2017). Under the resulting quasi-random representation of the sample selection, design-based methods may be employed, for example, in the construction of survey weights. In the case of non-probability sampling, Elliott and Valliant (2017) give particular attention to a case where an additional probability sample from the population is available for use in determining *pseudo-weights* for the nonprobability sample.

The second broad method is referred to as a *superpopulation model* approach by Elliott and Valliant (2017) and involves the model-based prediction approach outlined in Section 2.1.2. This depends critically on the auxiliary information available. The aim is to find auxiliary information so that, once conditioned upon, sample selection is noninformative. Moreover, the auxiliary variables are used to improve precision via regression-type models. A key concern with both of these approaches is the potential for bias as a result of departures from the modeling assumptions employed.

6. VARIANCE ESTIMATION

Interval estimation based on complex survey data is typically conducted by appealing to asymptotic normality. Thus, a conventional $100(1-\alpha)\%$ confidence interval for $\overline{Y}_U$ based on the HT estimator $\overline{y}_{\text{HT}}$ takes the form $(\overline{y}_{\text{HT}} - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\overline{y}_{\text{HT}})}, \overline{y}_{\text{HT}} + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\overline{y}_{\text{HT}})})$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $\widehat{\text{var}}(\overline{y}_{\text{HT}})$ is a consistent estimator of the variance of $\overline{y}_{\text{HT}}$. The basic idea is that the sampling distribution of the point estimator can be approximated by a normal distribution for large samples. The asymptotic theory used to justify the validity of such an interval needs to take account of the complexity of the design and is discussed by Breidt and Opsomer (2017). Given such an approach to interval estimation, the key tasks are to identify a suitable point estimator for a specified parameter of interest and a suitable estimator of the variance of this point estimator.

We have already seen that design consistent point estimation is often available using survey weighting for a range of designs. Given the complexity of weight construction, as discussed, for example, in Haziza and Beaumont (2017), it is common to separate this task, as a single exercise often undertaken by the agency conducting the survey, from the task of incorporating weights in estimation, undertaken by a wide range of analysts. There may be further reasons for such separation of tasks. For example, confidentiality considerations may impose restrictions on what information can be supplied to the analyst.

Given that variance estimation is arguably an even more complex challenge than weighting, there can be a similar rationale for task separation: focussing first on adding design information to the data file which can be used at a second stage by analysts to estimate variances for estimators of multiple targets. In principle, one could imagine adding to the data file joint probabilities of selection π_{kl} for all pairs of sample units so that the variance estimator in (4) could be computed. This is rarely done, however, in particular because all π_{kl} may simply be unavailable, as noted in Section 2.1.1. Indeed, the π_{kl} may not even be computable for many commonly used methods for selecting clusters in multi-stage sampling. Likewise, for model-based inference, the full design information relating to sample selection will often also be unavailable, perhaps for confidentiality reasons. Instead, there are certain "standard" kinds of information made available in survey data files to enable variance estimation to be conducted.

One approach is to approximate the actual design by a similar one that samples with replacement, rather than without. This is a common approach with stratified multistage designs in which the PSUs are selected with unequal probabilities within strata. For this

design, just the strata identifiers, PSU identifiers and weights provide sufficient design information for constructing consistent variance estimators. Valliant, Dever and Kreuter (2013), Chapter 15, describe how this may be done for linear estimators such as the HT estimator. For more complex nonlinear estimators, the method of *linearization* (more commonly referred to as the delta method in mainstream statistics) may be employed. This approach depends on the nature of the estimator, but is implemented in most statistical survey software. Other approximations, free of joint selection probabilities, and usable for design-based variance estimation, are reviewed by Berger and Tillé (2009).

Another broad approach is replication variance estimation. The bootstrap and jackknife methods are perhaps the most well-known examples; each of these are standard techniques in statistics; see Shao and Tu (2012) for a book length treatment. Suppose that survey weights, denoted $w_k, k \in S$, are used to estimate a parameter, $\overline{y}_U$, say, via a weighted estimator $\widehat{\overline{y}} = \sum_{k\in S} w_k y_k$. These weights may be simple design weights, or include post-stratification, nonresponse, etc., adjustments. Sets of L replicate weights $w_k^{(l)}, k \in S$ for $l = 1, \ldots, L$, are constructed and the variance estimator takes the form:

$$\sum_{l=1}^{L} c_l \big(\widehat{\overline{y}}^{(l)} - \widehat{\overline{y}}\big)^2,$$

where $\widehat{\overline{y}}^{(l)} = \sum_{k\in S} w_k^{(l)} y_k$. For suitable constants c_l, depending on the replication method, such replication weights can be constructed for a range of designs to achieve consistent variance estimation. The data file, released by the survey agency, now contains an additional L fields corresponding to the replicate weights, alongside the basic weights w_k.

For both the bootstrap and the jackknife, the replication weights $w_k^{(l)}, k \in S$ contain zeros, either from systematic deletion (the jackknife) or as a result of random subsampling (the bootstrap). The implementation of each of these techniques requires care since one must acknowledge the complex design. For example, under multistage sampling one may remove a complete PSU, which preserves the dependence structure of responses in the same cluster, and the weights are adjusted so as to preserve the sum of the weights. A further replication technique is balanced repeated replication (BRR). Under BRR, symmetries within the design are exploited to produce variance estimates from partially independent splits of the data. The bootstrap, jackknife and BRR techniques are discussed more fully in Rust and Rao (1996).

Another approach to variance estimation in complex designs is to adopt a model-based approach and to accommodate the population complexity in the model. For example, the induced dependence between units in a clustered population may be acknowledged in a model-based approach using mixed models, an approach that was championed by Scott and Smith (1969). In general, sandwich estimation is often utilized when adopting a model-based approach (Pfeffermann et al., 1998, Rabe-Hesketh and Skrondal, 2006). The underlying idea behind sandwich estimation is the empirical construction of variances under a (usually) simple working model. Sandwich estimation produces consistent standard error estimates under reduced assumptions when compared with a model-based approach, and is robust to the assumed variance model.

7. CONCLUDING REMARKS

In this short paper, we have given an overview of the design- and model-based approaches to inference for complex survey data. There are many important and emerging topics that we have not touched upon. For example, combining different sources of data is being increasingly carried out, and Lohr and Raghunathan (2017) review this endeavor.

A particular reason for the growing interest in combining data sources is the increased availability of "big data" sources, as noted in Section 1. There is also huge current interest in a landslide of associated new data analysis techniques, which often have their origins in machine learning. However, while many of these methods appear intoxicating, they must be carefully assessed to see whether they will provide valid inferences in the face of multiple sources of noncoverage and selection. If such aspects are ignored, there is a genuine possibility that big data analyses will produce really big inferential train wrecks.

ACKNOWLEDGEMENTS

Jon Wakefield was supported by award R01CA095994, from the National Institute of Health.

REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling, part I. In *Foundations of Statistical Inference* (*Proc. Sympos., Univ. Waterloo, Waterloo, Ont.*, 1970) 203–242. Holt, Rinehart and Winston, Toronto. MR0423625

BERGER, Y. G. and TILLÉ, Y. (2009). Sampling with unequal probabilities. In *Handbook of Statistics*, *Vol*. 29*A*, *Sample Surveys*: *Design*, *Methods and Applications* (D. Pfeffermann and C. R. Rao, eds.) 39–54. North-Holland, Amsterdam. MR2654632

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int*. *Stat*. *Rev*. **51** 279–292. MR0731144

BREIDT, J. and OPSOMER, J. (2017). Model-assisted survey estimation with modern prediction techniques. *Statist*. *Sci*. **32** 190–205.

BREWER, K. (2002). *Combined Survey Sampling Inference*: *Weighing Basu's Elephants*. Arnold, London.

CHAMBERS, R. L. and CLARK, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford Univ. Press, Oxford. MR3186498

CHEN, Q., ELLIOTT, M. R., HAZIZA, D., YANG, Y., GHOSH, M., LITTLE, R., SEDRANSK, J. and THOMPSON, M. (2017). Approaches to improving survey-weighted estimates. *Statist*. *Sci*. **32** 227–248.

COX, D. R. (2006). *Principles of Statistical Inference*. Cambridge Univ. Press, Cambridge. MR2278763

ELLIOTT, M. and VALLIANT, R. (2017). Inference for nonprobability samples. *Statist*. *Sci*. **32** 249–264.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist*. *Sci*. **22** 153–164. MR2408951

GRAUBARD, B. I. and KORN, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statist*. *Sci*. **17** 73–96. MR1910075

HÁJEK, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, part I', by D. Basu. In *Foundations of Statistical Inference* (*Proc*. *Sympos*., *Univ*. *Waterloo*, *Waterloo*, *Ont*., 1970). Holt, Rinehart and Winston, Toronto.

HAZIZA, D. and BEAUMONT, J.-F. (2017). Construction of weights in surveys: A review. *Statist*. *Sci*. **32** 206–226.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J*. *Amer*. *Statist*. *Assoc*. **47** 663–685. MR0053460

JAPEC, L., KREUTER, F., BERG, M., BIEMER, P., DECKER, P., LAMPE, C., LANE, J., O'NEIL, C. and ASHER, A. (2015). Big data in survey research: Aapor task force report. *Public Opin*. *Q*. **79** 839–880.

LITTLE, R. J. (2013). Calibrated Bayes, an alternative inferential paradigm for official statistics (with discussion). *J*. *Off*. *Stat*. **28** 309–372.

LOHR, S. L. (2010). *Sampling*: *Design and Analysis*, 2nd ed. Brooks/Cole Cengage Learning, Boston, MA.

LOHR, S. and RAGHUNATHAN, T. (2017). Combining survey data with other data sources. *Statist*. *Sci*. **32** 293–312.

LUMLEY, T. (2010). *Complex Surveys*: *A Guide to Analysis Using R*. Wiley, Hoboken, NJ.

LUMLEY, T. and SCOTT, A. (2017). Fitting regression models to survey data. *Statist*. *Sci*. **32** 265–278.

PFEFFERMANN, D., SKINNER, C. J., HOLMES, D. J., GOLDSTEIN, H. and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J*. *R*. *Stat*. *Soc*. *Ser*. *B* **60** 23–40. MR1625668

RABE-HESKETH, S. and SKRONDAL, A. (2006). Multilevel modelling of complex survey data. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *A* **169** 805–827. MR2291345

RUST, K. F. and RAO, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Stat*. *Methods Med*. *Res*. **5** 283–310.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SCHONLAU, M. and COUPER, M. (2017). Options for conducting web surveys. *Statist*. *Sci*. **32** 279–292.

SCOTT, A. and SMITH, T. M. F. (1969). Estimation in multi-stage surveys. *J*. *Amer*. *Statist*. *Assoc*. **64** 830–840.

SEAMAN, S. R. and WHITE, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Stat*. *Methods Med*. *Res*. **22** 278–295.

SHAO, J. and TU, D. (2012). *The Jackknife and Bootstrap*. Springer, New York. MR1351010

SKINNER, C. J. (2003). Introduction to part b. In *Analysis of Survey Data* (R. L. Chamber and C. J. Skinner, eds.) 75–84. Wiley, Chichester. MR1978845

SKINNER, C. J., HOLT, D. and SMITH, T. M. F., eds. (1989). *Analysis of Complex Surveys*. Wiley, Chichester. MR1049386

TILLÉ, Y. and WILHELM, M. (2017). Probability sampling designs; principles for the choice of design and balancing. *Statist*. *Sci*. **32** 176–189.

VALLIANT, R., DEVER, J. A. and KREUTER, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, Berlin. MR3088726

VALLIANT, R., DORFMAN, A. H. and ROYALL, R. M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. Wiley-Interscience, New York. MR1784794