University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2003   *Paper* 4

To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling

Rod Little[*]

[*]University of Michigan, rlittle@umich.edu

To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling

Rod Little

Abstract

Finite population sampling is perhaps the only area of statistics where the primary mode of analysis is based on the randomization distribution, rather than on statistical models for the measured variables. This article reviews the debate between design and model-based inference. The basic features of the two approaches are illustrated using the case of inference about the mean from stratified random samples. Strengths and weakness of design-based and model-based inference for surveys are discussed. It is suggested that models that take into account the sample design and make weak parametric assumptions can produce reliable and efficient inferences in surveys settings. These ideas are illustrated using the problem of inference from unequal probability samples. A model-based regression analysis that leads to a combination of design-based and model-based weighting is described.

To model or not to model? Competing modes of

inference for finite population sampling

Roderick J. Little

Department of Biostatistics

University of Michigan

Abstract

Finite population sampling is perhaps the only area of statistics where the primary mode of analysis is based on the randomization distribution, rather than on statistical models for the measured variables. This article reviews the debate between design and model-based inference. The basic features of the two approaches are illustrated using the case of inference about the mean from stratified random samples. Strengths and weakness of design-based and model-based inference for surveys are discussed. It is suggested that models that take into account the sample design and make weak parametric assumptions can produce reliable and efficient inferences in surveys settings. These ideas are illustrated using the problem of inference from unequal probability samples. A model-based regression analysis that leads to a combination of design-based and model-based weighting is described.

Keywords: Bayesian methods; design-based inference; sampling weights; regression, robustness; survey sampling

1. Introduction

Scientific survey sampling, as represented by Neyman's (1934) classic paper and subsequent developments (e.g. Hansen and Hurwitz 1943; Mahalanobis 1946) is one of the greatest contributions of statistics to science. It provides the remarkable ability to obtain useful inferences about large populations from modest samples, with measurable uncertainty. Extensions of simple random sampling to stratified multistage sampling greatly extend the reach of scientific sampling in the real world, and form the backbone of data collection in science and government.

2

The key role of random sampling for *data collection* is not at issue in this article. The question concerns the role of the randomization distribution in the statistical *analysis* of random survey samples. Survey sampling is perhaps unique in being the only area of current statistical activity where inferences are primarily based on the randomization distribution rather than on statistical models for the survey outcomes. It is an area where the debate between randomization-based and model-based inference is most sharply drawn (e.g. Smith 1976, 1994; Kish, 1995). These philosophical differences in the analysis of survey data arise early in the study of statistics, in the form of the role of weights in multiple regression. The following example describes the issue.

Example 1. Weights in regression. The basic fitting algorithm for standard forms of normal linear regression is ordinary least squares (OLS). In an early course on statistical methods, we learn that OLS is based on a model that assumes that the residual variance is constant for all values of the covariates. If the variance of the residual for unit i is s^2/u_i for some known constant u_i, then better inferences are obtained by weighted least squares, with unit i weighted proportional to u_i. This form of weighting is model-based, since the linear regression model for the outcome (say *Y*) has been modified to incorporate a non-constant residual variance.

A quite different form of weighting arises in survey sampling, based on the selection probabilities. If unit i is sampled with selection probability p_i, then the survey sampler replaces OLS by weighted least squares, weighting the contribution of unit i to the least squares equations by $w_i \propto 1/p_i$, the inverse of the probability of selection. This form of weighting is design-based, with p_i relating to the selection of units: since unit i

3

"represents" $1/p_i$ units of the population, it receives a weight proportional to $1/p_i$ in the regression.

Both forms of weighting seem plausible, but they are not necessarily the same. So if they are different, which is correct? The central role of the mode of inference to this question is clear, since the modeler's distribution of Y seems to lead to weighting by u_i and the randomization distribution leads to weighting by w_i. The role of sampling weights in regression has been extensively debated in the literature; see for example, Konijn (1962), Brewer and Mellor (1973), Dumouchel and Duncan (1983), Smith (1988), Little (1991), Pfeffermann (1993), Korn and Graubard (1999). In the concluding section I discuss another regression estimator that weights cases by the product of the design and model weights. I show how this estimator can be justified from both frequentist and model-based perspectives, thus harmonizing the two approaches to inference.

Many survey statisticians adopt both design and model-based philosophies of statistical analysis, according to the context. For example, descriptive inference about finite population quantities based on large probability samples are carried out using design-based methods, but models are used for problems where this does approach not work, such as nonresponse or small area estimation. This pragmatic approach has increased in popularity since battles over the "foundations of survey inference" in the 1980's subsided. While the application of statistics to real data requires pragmatism, I have always felt the need for an unambiguous underlying theory. Just as mathematicians do not tolerate two competing theories of differential calculus, we should not be happy with two competing statistical theories that can lead to different solutions. Thus to avoid

4

"inferential schizophrenia", I have always sought to reconcile the best aspects of survey analysis within a single statistical theory, namely, Bayesian modeling.

Advocating Bayes for sample survey inference is "swimming upstream", since its subjectivist basis is anathema to many survey statisticians, who do not like modeling assumptions. But Bayesian methods run the gamut of subjectivity, and can be as "objective" as any frequentist method when necessary; indeed many frequentist answers can be replicated from a Bayesian perspective.

This article reviews some of the issues that inform the design-based and model-based debate concerning the analysis of sample survey data. Section 2 outlines the basic features of design and model-based survey inference. Section 3 describes strengths and weaknesses of design-based inference, and Section 4 considers "model-assisted" survey inference, which captures some of the positive features of models within the design-based paradigm. Section 5 discusses the modeling approach to survey inference, with particular reference to the issue of survey weighting raised in Example 1. Section 6 presents some conclusions, and speculates on possible future trends in sample survey analysis.

2. A brief review of design and model-based inference

The design-based approach to survey inference is described in many texts (e.g. Hansen, Hurwitz and Madow 1953, Kish 1965, Cochran 1977). The following description is not completely general, but captures the main features. For a population with N units, let $Y = (y_1,..., y_N)$ where y_i is the set of survey variables for unit i, and let $I = (I_1,...,I_N)$ denote the set of *inclusion indicator variables,* where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included. Design-based inference is based on

5

the distribution of *I*, with the survey variables *Y* treated as fixed quantities. For inference about a finite population quantity $Q = Q(Y)$ it involves the following steps:

(a) the choice of an estimator $\hat{q} = \hat{q}(Y_{inc}, I)$, a function of the observed part Y_{inc} of *Y*, that is unbiased or approximately unbiased for *Q* with respect to the distribution *I*. I like writing the inclusion indicators *I* as an explicit argument of $\hat{q}$ to emphasize that $\hat{q}$ is a random variable as a function of *I*, not Y_{inc}, which are fixed quantities.

(b) the choice of a variance estimator $\hat{v} = \hat{v}(Y_{inc}, I)$ that is unbiased or approximately unbiased for the variance of $\hat{q}$ with respect to the distribution of *I*.

Inferences are then generally based on normal large sample approximations. For example, a 95% confidence interval for *Q* is $\hat{q} \pm 1.96\sqrt{\hat{v}}$.

Example 2. Design-based inference for the mean from a stratified random sample.

To illustrate the above process, consider the simple case of estimation of a finite population mean $\overline{Y}$ from a stratified random sample. Suppose the population is divided into *J* strata, and let N_j be the known population count in stratum *j* and $\overline{Y}_j$ the unknown population mean in stratum *j*. The quantity of interest is $Q = \overline{Y} = \sum_{j=1}^{J} P_j \overline{Y}_j$, where

$P_j = N_j / N$ is the proportion of the population in stratum *j*. We assume that a random sample of size n_j of the N_j units are sampled in stratum *j*, and let $\{y_{ji}, i = 1,...,n_j\}$ denote the set of sampled *Y*-values in stratum *j*. Then $Y_{inc} = \{y_{ji}, j = 1,...,J; i = 1,...,n_j\}$.

Stratified random sampling has the property that all the possible samples of size n_j in stratum *j* have the same probability of being selected. Formally:

6

$$\Pr(I_{ji} = 1) = \left[\binom{N_j}{n_j} \right]^{-1}, \text{ if } \sum_{i=1}^{N_j} I_{ji} = n_j, \text{ and 0 otherwise}.$$

The usual estimator of $\bar{Y}$ in this setting is the stratified mean

$$\hat{q} = \bar{y}_{st} \equiv \sum_{j=1}^{J} P_j \bar{y}_j, \tag{1}$$

where $\bar{y}_j$ is the sample mean in stratum j. The estimator (1) is also a weighted mean of

the sampled units, where units in stratum j are weighted by the inverse of their selection

probability $p_j = n_j / N_j$. An attractive feature of stratified sampling is that the selection

probabilities can vary across strata, giving rise to the design weights discussed in

Example 1. The estimated variance of the stratified mean is

$$\hat{v}_{st} = \sum_{j=1}^{J} P_j^2 s_j^2 (1/n_j - 1/N_j), \tag{2}$$

where s_j^2 is the sample variance in stratum j. The quantities $\bar{y}_{st}$ and $\hat{n}_{st}$ are the basis of

95% confidence intervals of the form $\bar{y}_{st} \pm 1.96\sqrt{\hat{v}_{st}}$ for $\bar{Y}$, and tests for null values of

the population mean $\bar{Y}$.

The model-based approach to survey sampling inference requires a model for the

survey outcomes Y, which is then used to predict the non-sampled values of the

population, and hence finite population quantities Q. There are two major variants:

superpopulation modeling and Bayesian modeling. In superpopulation modeling (e.g.

Royall 1970; Thompson 1988; Valliant, Dorfman, and Royall 2000), the population

values of Y are assumed to be a random sample from a "superpopulation", and assigned a

probability distribution $p(Y | q)$ indexed by fixed parameters q. Inferences are based on

the joint distribution of Y and I.

7

Bayesian survey inference (Ericson 1969, 1988; Basu 1971; Scott 1977; Binder 1982; Rubin 1983, 1987; Ghosh and Meeden 1997) requires the specification of a prior distribution $p(Y)$ for the population values. Inferences for finite population quantities $Q(Y)$ are then based on the posterior predictive distribution $p(Y_{exc} | Y_{inc})$ of the non-sampled values (say Y_{exc}) of Y, given the sampled values Y_{inc}. The specification of the prior distribution $p(Y)$ seems a formidable task, but is often achieved via a parametric model $p(Y | \boldsymbol{q})$ indexed by parameters $\boldsymbol{q}$, combined with a prior distribution $p(\boldsymbol{q})$ for $\boldsymbol{q}$, that is:

$$p(Y) = \int p(Y | \boldsymbol{q}) \, p(\boldsymbol{q}) \, d\boldsymbol{q} \, .$$

The posterior predictive distribution of Y_{exc} is then

$$p(Y_{exc} | Y_{inc}) \propto \int p(Y_{exc} | Y_{inc}, \boldsymbol{q}) \, p(\boldsymbol{q} | Y_{inc}) d\boldsymbol{q} \, ,$$

where $p(\boldsymbol{q} | Y_{inc})$ is the posterior distribution of the parameters, computed via Bayes' Theorem:

$$p(\boldsymbol{q} | Y_{inc}) = p(\boldsymbol{q}) p(Y_{inc} | \boldsymbol{q}) / \, p(Y_{inc}) \, .$$

Here $p(\boldsymbol{q})$ is the prior distribution, $p(Y_{inc} | \boldsymbol{q})$ is the likelihood function, viewed as a function of $\boldsymbol{q}$, and $p(Y_{inc})$ is a normalizing constant. This posterior distribution induces a posterior distribution $p(Q | Y_{inc})$ for finite population quantities $Q(Y)$.

The specification of $p(Y | \boldsymbol{q})$ in this Bayesian formulation is the same as in parametric superpopulation modeling, and in large samples the likelihood based on this distribution dominates the contribution from the prior for $\boldsymbol{q}$. As a result, large-sample inferences from the superpopulation modeling and Bayesian approaches are often similar.

8

However, in my view the Bayesian approach is conceptually more straightforward and has some advantages for small samples, as illustrated in the next example.

The model formulations described thus far do not involve the distribution for *I*, basing inferences on the distribution of *Y* alone. This is justified when the sampling mechanism is "unconfounded" or "noninformative", as when the distribution of *I* given *Y* does not depend on the values of *Y* (Rubin 1987, Chambers 2003). This is indeed the case with probability sampling, but is not necessarily the case with other less well-controlled forms of sampling, such as quota sampling. If the sampling mechanism is confounded, then model inferences must be based on a model for the joint distribution of *I* and *Y*, rather than simply a model for the marginal distribution of *Y*, and formulating an acceptable model for confounded sampling mechanisms is problematic. A key motivation for probability sampling from the modeling perspective is that it avoids the need to specify a model for the sampling mechanism, even though the sampling distribution is not the basis for inference. From the Bayesian perspective, random sampling provides a justification for assumptions of exchangeability of the sampled units (De Finetti 1990) that underpin i.i.d. models, such as that discussed in the next example for the case of stratified sampling.

Example 3. Model-based inference for the mean from a stratified random sample.
Sensible parametric models for stratified samples need to reflect stratum differences by assigning distinct parameters to the distribution of *Y* in each stratum. (The reason is explained in Example 8 below). Let y_{ji} denote the value of *Y* for unit *i* in stratum *j*. A common baseline model for continuous outcomes assumes that y_{ji} is normal with mean

9

$\boldsymbol{m}_j$ and variance $\boldsymbol{s}_j^2$. A simple Bayesian specification in the absence of strong prior

knowledge adds a noninformative prior for the parameters $\{\boldsymbol{m}_j, \boldsymbol{s}_j^2\}$, yielding the model:

$$
\begin{aligned}
&p(y_{ji} \mid z_{ji} = j, \boldsymbol{q}) \sim_{iid} G(\boldsymbol{m}_j, \boldsymbol{s}_j^2); \boldsymbol{q} = \{\boldsymbol{m}_j, \boldsymbol{s}_j^2\} \\
&p(\boldsymbol{m}_j, \log \boldsymbol{s}_j^2) = const.,
\end{aligned}
\tag{3}
$$

where $G(\boldsymbol{m}_j, \boldsymbol{s}_j^2)$ denotes the normal (Gaussian) distribution with mean $\boldsymbol{m}_j$ and variance

$\boldsymbol{s}_j^2$. With known variances $\{\boldsymbol{s}_j^2\}$, standard Bayesian calculations for this model yields

the posterior distribution of $\bar{Y}$ given Y_{inc}, I and $\{\boldsymbol{s}_j^2\}$ as normal with mean

$$
\begin{aligned}
&E(\bar{Y} \mid Y_{inc}, I, \{\boldsymbol{s}_j^2\}) = \bar{y}_{st} = \sum_{j=1}^{J} P_j \bar{y}_j \\
&Var(\bar{Y} \mid Y_{inc}, I, \{\boldsymbol{s}_j^2\}) = v_{st} = \sum_{j=1}^{J} P_j^2 \boldsymbol{s}_j^2 (1/n_j - 1/N_j).
\end{aligned}
\tag{4}
$$

The posterior mean is the stratified mean (1) from design-based inference. When $\{\boldsymbol{s}_j^2\}$

are replaced by estimates $\{s_j^2\}$, the posterior variance equals the design-based variance

(2). This substitution is justified asymptotically. Thus in large samples, the posterior

distribution of $\bar{Y}$ yields a 95% posterior probability interval $\bar{y}_{st} \pm 1.96 \hat{v}_{st}$ that is the same

as the design-based 95% confidence interval in Example 2. The two approaches yield the

same interval estimate, although the Bayesian posterior probability interval has the direct

interpretation as a probability statement for the unknown population mean, rather than as

a confidence interval.

The full Bayesian analysis under (3) propagates the uncertainty in estimating the

variances $\{\boldsymbol{s}_j^2\}$ by integrating them out of the posterior distribution of $\bar{Y}$ given Y_{inc}, I

and $\{\boldsymbol{s}_j^2\}$ over the posterior distribution of $\{\boldsymbol{s}_j^2\}$ given Y_{inc}, I. The posterior distribution

of $\boldsymbol{s}_j^2 / \{(n_j - 1)s_j^2\}$ is easily shown to be inverse chi-squared with $n_j - 1$ degrees of

freedom, independently for $j = 1, \ldots, J$. Integrating over these posterior distributions yields the posterior distribution of $\overline{Y}$ given Y_{inc}, I as a mixture of t distributions. This posterior distribution has a complicated density, but draws from it are readily computed by (a) drawing $\tilde{s}_j^2 = (n_j - 1)s_j^2 / c_j$ where c_j is chi-squared with $n_j - 1$ degrees of freedom, and (b) drawing $\overline{Y}$ from a normal distribution with mean $\overline{y}_{st}$ and variance $\tilde{v}_{st}$, where the variances s_j^2 in v_{st} are replaced by their drawn values $\tilde{s}_j^2$. These draws can then be used to approximate the posterior distribution to any desired degree of accuracy. Note that integrating over the posterior distribution of $\{s_j^2\}$ rather than simply plugging in estimates yields a useful small-sample correction not readily available from design-based and superpopulation approaches.

3. Strengths and weaknesses of design-based inference

The design-based approach to survey inference has a number of strengths that make it popular with practitioners. It automatically takes into account features of the survey design, and it provides reliable inferences in large samples, without the need for strong modeling assumptions. On the other hand it is essentially asymptotic, and hence yields limited guidance for small-sample adjustments. Unlike models, which lead to efficient inferences based on likelihood or Bayesian principles, the design-based approach is not prescriptive for the choice of estimator. It lacks a theory for optimal estimation (Godambe 1955), and estimates from the approach are potentially inefficient. Consider the following important example.

11

Example 4. The Horvitz-Thompson estimator. I have noted that the stratified mean weights sampled units by the inverse of their probability of selection. The Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952) applies this idea more generally. Consider inference about the population total

$$Q(Y) = T \equiv Y_1 + ... + Y_N \,,$$

and any sample design with positive inclusion probability $\boldsymbol{p}_i = E(I_i \mid Y) > 0$ for unit i, $i = 1,\ldots,N$. The HT estimator is then

$$\hat{t}_{HT} = \sum_{i \, \text{sampled}} Y_i / \boldsymbol{p}_i = \sum_{i=1}^{N} I_i Y_i / \boldsymbol{p}_i \,, \tag{5}$$

and is design unbiased for T, since

$$E(\hat{t}_{HT} \mid Y) = \sum_{i=1}^{N} E(I_i \mid Y) Y_i / \boldsymbol{p}_i = \sum_{i=1}^{N} \boldsymbol{p}_i Y_i / \boldsymbol{p}_i = \sum_{i=1}^{N} Y_i \,.$$

The unbiasedness of (5) under very mild conditions conveys robustness to modeling assumptions, and makes it a mainstay of the design-based approach. But (5) has two major deficiencies. First, the choice of variance estimator is problematic for some probability designs (e.g. systematic sampling). Second, the HT estimator can have a high variance, for example, when an outlier in the sample has a low selection probability, and hence receives a large weight. Basu's (1971) famous circus elephant example provides an amusing, if extreme example:

> "The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take

12

50y (where y is the present weight of Sambo) as an estimate of the total weight $Y = Y_1 + Y_2 + ...Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive sampling plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why? The estimate ought to be 50y of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators."What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is 100y/99 and not 50y." "And, how would you have estimated Y," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been 4900y, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!)"

The practical bent of survey samplers is illustrated by the fact that Basu (a Bayesian) makes fun of the frequentist position by placing it in the domain of "mathematical statistics". On the other side, Leslie Kish, an avid design-based advocate, similarly criticizes mathematical statisticians for focussing on i.i.d. models that fail to account for the complex sample design (Kish 1995, Section 9).

Randomization inference suffers from ambiguity about the appropriate reference distribution in certain problems. This issue arises in sample survey settings, as in the following example:

13
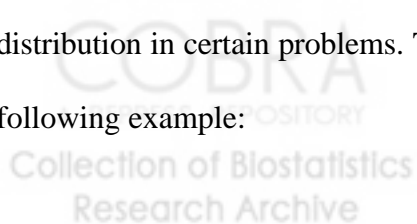
Example 5. Post-stratification. Another form of weighting arises in design-based inference with post-stratification. Sometimes, the population distribution is known (from external data such as a Census) for a variable that is not observed for all population units prior to sampling, and hence cannot be used as a stratifier. In this case, it is still possible to use the distribution to adjust estimates of the outcome in the analysis using the technique known as post-stratification. Suppose the quantity of interest is $Q = \bar{Y} = \sum_{j=1}^{J} P_j \bar{Y}_j$, where $P_j = N_j / N$ is now the proportion of the population in post-stratum j. We assume that a random sample of size n is selected from the population, and n_j of the N_j units in post-stratum j are included in the sample; unlike stratification, the distribution of $\{n_j\}$ is now not under the control of the sample, and varies from sample to sample. The usual estimator of $\bar{Y}$ is then the post-stratified mean

$$\hat{q} = \bar{y}_{ps} \equiv \sum_{j=1}^{J} P_j \bar{y}_j , \tag{6}$$

where $\bar{y}_j$ is the sample mean in post-stratum j. The estimator (6) has the same form as the stratified mean (1), and is also a weighted mean of the sampled units, where units in post-stratum j are given the post-stratification weight N_j / n_j. More generally, in complex sample designs, a post-stratification weight is often applied as a multiplicative factor, after weighting for sample selection and nonresponse.

Since $\bar{y}_{ps}$ has the same form as $\bar{y}_{st}$, one might expect design-based inferences to be analogous. However, the design-based variance of $\bar{y}_{ps}$ is changed by the fact that $\{n_j\}$ are now random functions of the sampling distribution I. In fact, in repeated sampling of I, there is a non-zero probability that $n_j = 0$ for some j, in which case $\bar{y}_{ps}$ is undefined!

14

Hence the design-based variance of $\bar{y}_{ps}$ is undefined, or maybe infinite! The usual resolution of this problem is to condition on $\{n_j\}$ observed in the realized sample, on the grounds that these counts are a form of ancillary statistic, and modify the post-strata to ensure that $\{n_j\}$ are all greater than zero. The fact that $\bar{y}_{ps}$ is design-unbiased conditionally on $\{n_j\}$ might be construed as a form of ancillarity, but a formal frequentist theory on when and how to choose ancillary statistics is lacking for classical parametric models (Cox and Hinkley 1974, Example 2.28), and is even less developed in the finite sampling setting. Also, the sample mean $\bar{y} = \sum_{j=1}^{n} n_j \bar{y}_j / n$, the standard estimator in the absence of the post-stratum counts, is not design-unbiased conditionally on $\{n_j\}$; it seems awkward to vary the reference distribution according to whether the post-stratified or unweighted mean is used to estimate $\bar{Y}$.

Conditioning on $\{n_j\}$ leads to the variance $v_{ps} = \sum_{j=1}^{J} P_j^2 S_j^2 / n_j$, where S_j^2 is the population variance of Y in post-stratum j, ignoring finite population corrections. A practical issue stemming from the lack of control of $\{n_j\}$ is that we may be unlucky and draw a sample where S_j^2 is large and n_j is small in one or more post-strata, yielding a large $\boldsymbol{n}_{ps}$, a practical illustration of the problem caricatured in Example 4. From a prediction point of view, the problem lies in the lack of information with which to estimate $\bar{y}_j$ in these sparse cells. A method is needed for "borrowing strength" from Y-values in other post-strata. In practice, this problem is often mitigated by combining post-strata with small counts with neighboring post-strata. A more systematic approach to borrowing strength is to base it on a model for Y, as discussed in Example 9 below.

15

Another limitation of design-based inference is that it is strictly inapplicable to situations where the randomization distribution is corrupted by non-sampling errors, such as nonresponse or measurement errors; modeling assumptions are needed to address these problems. Kalton (2002) reviews these limitations of design-based inference.

4. Model-assisted design-based inference

Superpopulation models are not the basis for inference in the design-based approach, but they can be useful to motivate the choice of estimator; in particular many of the classical estimators for incorporating covariate information, such as the ratio estimator or the regression estimator (e.g. Cochran 1977), can be motivated as arising from linear superpopulation models. The next example views the HT estimator from this perspective.

Example 6. A model for the Horvitz-Thompson estimator (Example 4 continued).
The HT estimator can be regarded as a model-based estimator for the following linear model relating y_i to $\boldsymbol{p}_i$:

$$y_i = \boldsymbol{b}\boldsymbol{p}_i + \boldsymbol{p}_i\boldsymbol{e}_i, \tag{7}$$

or equivalently,

$$z_i = y_i/\boldsymbol{p}_i = \boldsymbol{b} + \boldsymbol{e}_i, \tag{8}$$

where $\boldsymbol{e}_i$ in Eqs. (7) and (8) are assumed to be i.i.d. normally distributed with mean zero and variance $\boldsymbol{s}^2$. Models (7) or (8) lead to $\hat{\boldsymbol{b}} = n^{-1}\sum_{i\in S} y_i/\boldsymbol{p}_i = t_{HT}/n$ where n is the

16

sample size. The corresponding prediction for unit i is $\hat{y}_i = \hat{\boldsymbol{b}}\boldsymbol{p}_i$, and the prediction estimator of the total is thus

$$\hat{T}_{\text{pred}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \in S} (y_i - \hat{y}_i) = \hat{t}_{HT} + \sum_{i \in S} (y_i - \hat{y}_i),$$

which differs from the HT estimator by a quantity that tends to zero with the sampling fraction n/N. This analysis suggests that the HT estimator is likely to be good estimator when (7) or (8) is a good description of the population, and it may be inefficient when it is not. A formal explanation for the poor properties of the HT estimator of the elephants' total weight in Example 4 is that the model (7) is clearly inappropriate, given the way the weights are chosen.

Example 7. The Generalized Regression Estimator. In situations where the HT model is not reasonable, a model-assisted modification is to predict the non-sampled values using a more suitable model, and then apply the HT estimator to the residuals from that model. Specifically, the generalized regression estimator of T takes the form:

$$\hat{T}_{\text{gr}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \text{ sampled}} (y_i - \hat{y}_i)/\boldsymbol{p}_i, \qquad (9)$$

where $\hat{y}_i$ is the prediction from a linear regression model relating Y to the covariates. The second term on the right side of (9) conveys it with the useful property of design consistency (Brewer 1979, Isaki and Fuller 1982), which means informally that the estimator converges to the population quantity being estimated as the sample size increases, in a manner that maintains the features of the sample design. Design-based statisticians usually weight cases by the design weights w_i when computing this

17

regression, but the estimator (9) is also design consistent if the regression is variance weighted. For discussions of generalized regression estimator and alternatives, see for example Cassel, Särndal and Wretman (1977), Särndal, Swensson, and Wretman (1992).

Another general approach to design-based inference incorporate models by basing inference on "pseudo-likelihoods" that reflect survey design features (Binder, 1983; Godambe and Thompson, 1986). Suppose a superpopulation model is posited for the N population units of the form:

$$p(y \mid z, \boldsymbol{q}) = \prod_{i=1}^{N} p(y_i \mid z_i, \boldsymbol{q}), \tag{10}$$

which assumes independence across units. If the whole population were sampled, standard model-based inference would be based on the loglikelihood

$$\ell_{\text{pop}}(\boldsymbol{q} \mid y_{\text{inc}}, z) \propto \sum_{i=1}^{N} \log p(y_i \mid z_i, \boldsymbol{q}).$$

Under mild conditions the ML estimate would be obtained by solving the score equations obtained by differentiating the loglikelihood with respect to $\boldsymbol{q}$, that is

$$sc_{\text{pop}}(\boldsymbol{q}) = \sum_{i=1}^{N} \partial \log p(y_i \mid z_i, \boldsymbol{q}) / \partial \boldsymbol{q} = 0. \tag{11}$$

For any value of $\boldsymbol{q}$, $sc_{\text{pop}}(\boldsymbol{q})$ is a finite population quantity that can be estimated from the sample. The "pseudo-likelihood" approach estimates the score by a design-consistent estimator, and solves the resulting "estimated" score equation. For example one might apply HT weighting to (11), yielding the estimated score equation

$$sc_{HT}(\boldsymbol{q}) = \sum_{i=1}^{N} I_i \big(\partial \log p(y_i \mid z_i, \boldsymbol{q}) / \partial \boldsymbol{q} \big) / \boldsymbol{p}_i = 0. \tag{12}$$

18

In the special case of normal linear regression, maximizing $sc_{HT}(\boldsymbol{q})$ yields least squares estimates with sampled unit i weighted by the sampling weight $\boldsymbol{p}_i^{-1}$. Eq. (12) generalizes the HT estimator, but does not overcome its potential lack of efficiency noted above. The approach is not prescriptive about how to estimate the score, particularly in settings where the assumption of independent observations in Eq. (11) is not warranted, as in multistage sampling. Pfeffermann et al. (1998) discuss how this approach might be adapted to multilevel models, but their suggestions lack general guiding principles.

5. Model-based inference

I now turn to inferences based on superpopulation or Bayesian models. Some advantages of this approach are:

(1) it provides a unified approach to survey inference, aligned with mainline statistics approaches in other application areas such as econometrics.

(2) In large samples or for models with uninformative prior distributions, results can parallel those from design-based inference, as we have seen in the case of stratified sampling in Examples 2 and 3.

(3) The Bayesian approach is well equipped to handle complex design features such as clustering through random cluster models (Scott and Smith 1969), stratification through covariates that distinguish strata, nonresponse (Little 1982; Rubin 1987; Little and Rubin 2002) and response errors.

(4) The Bayesian approach may yield better inferences for small sample problems where exact frequentist solutions are not available, by propagating error in estimating parameters. For example, the posterior distribution of the mean for

inference from normal stratified samples in Example 3 is a mixture of t distributions that propagates uncertainty in estimating the stratum variances. On the other hand, the standard design-based inference based on the normal distribution assumes that the stratum variances are estimated without error from the sample.

(5) The Bayesian approach allows prior information to be incorporated, when appropriate; and

(6) The Bayesian approach avoids the ambiguities in the choice of reference distribution discussed in Example 5, and has useful features of coherency not shared by frequentist approaches, such as satisfying the likelihood principle.

(7) Likelihood-based approaches like Bayes or maximum likelihood have the property of large-sample efficiency, and hence match or outperform design-based inferences if the model is correctly specified.

The challenge with the modeling approach lies in the last phrase: how exactly to specify the model? All models are simplifications and hence subject to some degree of misspecification. The major weakness of model-based inference is that if the model is seriously misspecified it can yield inferences that are worse (and potentially much worse) than design-based inferences. The following example might serve as a design-based statistician's rejoinder to the "Basu elephant" disaster in Example 4:

Example 8: a non-robust model for disproportionate stratified sampling. In the setting of disproportionate stratified sampling (Example 3), models are needed that condition on stratum in order for the sample design to be unconfounded. Suppose a

20

normal model is posited that assumes the distribution of Y is the same for all strata, that is:

$$p(y_{ji} \mid z_{ji} = j, \boldsymbol{q}) \sim_{iid} G(\boldsymbol{m}, \boldsymbol{s}^2); \boldsymbol{q} = \{\boldsymbol{m}, \boldsymbol{s}^2\}$$
$$p(\boldsymbol{m}, \log \boldsymbol{s}^2) = const.$$
(13)

The posterior mean of $\overline{Y}$ under this model is the unweighted sample mean

$\overline{y} = \sum_{j=1}^{J} n_j \overline{y}_j / n$. This is the same as the stratified mean in equal probability samples,

but differs when the probabilities of selection vary across the strata. If the model (13)

were known to be true, as for example if the strata were created using random numbers,

then the unweighted mean is a better estimator than the stratified mean. However, in

practice strata are never created in this way, but rather are based on characteristics likely

to be related to the survey outcomes. If the sample size is large, even a slight

misspecification in (13) caused by minor differences in the distribution of Y between

strata can induce a bias in $\overline{y}$ that dominates mean squared error and corrupts confidence

coverage. Hansen, Madow, and Tepping (1983) show in a related example that the bias

can be serious even when diagnostic checks for differences between strata are negative.

Modelers have questioned Hansen et al.'s the choice of diagnostics (Valliant, Dorfman,

and Royall 2000), but my view is that a model such as (13) that ignores stratum effects is

too vulnerable to misspecification to be a reliable basis for inference, unless there are

convincing reasons to believe that stratum effects are not present. For more discussion of

the adverse effects of model misspecification on survey inference, see Kish and Frankel

(1974), Holt, Smith, and Winter (1980), and Pfeffermann and Holmes (1985).

    Inferential disasters can be avoided by selecting models that are attentive to

design features such as stratification and clustering. Since the design of the sample in a

21

passive observational study has no effect on the population values, in principle the choice of model should not be affected by the sample design. However, in practice all models are simplifications, and the features of the population that are important to include in the model *do* vary according to the choice of design. In particular, for inferences about a population mean in Example 8, it is important to model stratum differences when the sample is selected by disproportionate stratified sampling, but modeling these differences becomes unimportant when the sample is selected by simple random sampling. It is important to incorporate spatial correlation into the model when the sample design involves spatial clustering, but spatial correlation is not an important feature of a model for an unclustered sample. I think choosing a model that incorporates important design features is conceptually more satisfying than fixing a deficient model using the methods in Section 4.

One way of limiting the effects of model misspecification is to restrict attention to models that yield design-consistent estimates. This limitation is not as restrictive as it may seem; a number of strategies are discussed in Firth and Bennett (1998). In the context of surveys with non-constant inclusion probabilities, a key is to model differences in the distribution of outcomes across classes defined by differential probabilities of inclusion (Rubin 1985; Rizzo 1992). The following model leads to a number of interesting special cases. Let y_{ji} denote the outcome for unit *i* in inclusion class *j*, within which the inclusion probability is constant. Suppose for simplicity that the proportion of the population in inclusion class *j*, P_j, is known; in cases where it is unknown a supplemental model is needed to allow estimation of these proportions from the sample. Consider the mixed effects model:

$$[y_{ji} \mid \boldsymbol{m}_j, \boldsymbol{s}_j^2] \sim_{ind} G(\boldsymbol{m}_j, k_{1j}\boldsymbol{s}_j^2)$$
$$[\boldsymbol{m}_j \mid P_j, C_j, \boldsymbol{b}, \boldsymbol{t}^2] \sim_{ind} G(y_j^*, k_{2j}\boldsymbol{t}^2), \ y_j^* = f(P_j, C_j; \boldsymbol{b}) \qquad (14)$$
$$[\boldsymbol{b}, \log \boldsymbol{t}^2, \log \boldsymbol{s}_i^2] \sim const.$$

Here k_{1j} and k_{2j} are known constants that model heteroskedasticity, and $f(.)$ is a known

function of P_j and covariates C_j characterizing the inclusion classes, indexed by

unknown regression parameters $\boldsymbol{b}$. Two extreme forms of this model are noteworthy.

When $\boldsymbol{t}^2 = \infty$ we obtain a *fixed-effects* version of the model that estimates the mean in

each inclusion class j by the sample mean $\overline{y}_j$. The resulting estimate of the population

mean (ignoring finite population corrections) is $\sum_{j=1}^{J} P_j \overline{y}_j$, which is equivalent to the

design-weighted estimator. When $\boldsymbol{t}^2 = 0$, $\boldsymbol{m}_j = y_j^*$, and we obtain a *direct regression*

version of the model. The resulting estimate of the population mean (ignoring finite

population corrections) is $\sum_{j=1}^{J} P_j \hat{y}_j$, where $\hat{y}_j = g(P_j, C_j, \hat{\boldsymbol{b}})$ is the prediction of the

mean in inclusion class j from the regression model. Estimates from (14) with $0 < \boldsymbol{t}^2 < \infty$

shrink the sample mean from fixed-effects model towards the prediction from the

regression model. The degree of shrinkage goes to zero as the sample increases, which

implies that estimates from the model are design consistent. On the other hand the

regression feature allows borrowing of strength for the predicted means of small

inclusion classes. The next two examples concern special cases of model (14).

Example 9. A model for improving the stratified or post-stratified mean. Suppose the

inclusion classes are strata with differential inclusion probabilities $\{\boldsymbol{p}_j : j = 1,...J\}$, where
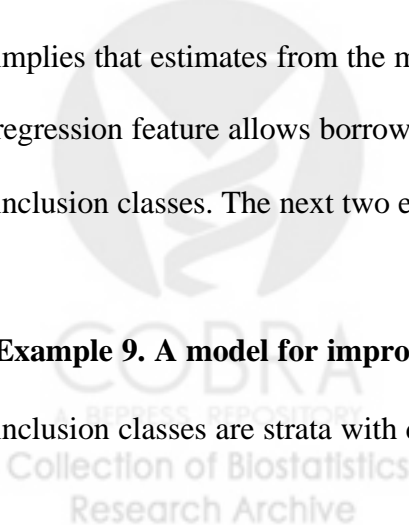
$\boldsymbol{p}_1 < \boldsymbol{p}_2 < ... < \boldsymbol{p}_J$, and consider the model (14) with $k_{1j} = k_{2j} = 1$ and $y_j^* = \boldsymbol{m}$, a constant.

A standard random-effects model analysis yields

$$E(\boldsymbol{m}| data, \{\boldsymbol{s}_j^2\}, \boldsymbol{t}^2) = \sum_{j=1}^{J} P_j \{\boldsymbol{l}_j \bar{y}_j + (1 - \boldsymbol{l}_j) \bar{y}_I\},$$

where $\boldsymbol{l}_j = n_j \boldsymbol{t}^2 / (n_j \boldsymbol{t}^2 + \boldsymbol{s}^2)$ and $\bar{y}_I = \sum_{j=1}^{J} n_j \boldsymbol{l}_j \bar{y}_j / \sum_{j=1}^{J} n_j \boldsymbol{l}_j$. This estimate shrinks the

(post)stratified mean $\sum_{j=1}^{J} P_j \bar{y}_j$ towards the unweighted mean, and yields a form of

empirically-based weight smoothing. In practice the variance components can be

estimated, or a fully Bayesian analysis carried out using the Gibbs' sampler (Gelfand et

al, 1990).

Better models adopt a more realistic regression structure. For example, Elliott and

Little (2000) shrink the (post)stratified means towards a smooth function of the selection

probabilities, determined by a spline function. This approach yields gains in precision

when the sample weights are variable, and is robust to model misspecification since the

form of the model is weak.

Example 10. A model for improving the HT estimator in PPS samples. In the case of

sampling with probability proportional to size, inclusion classes often contain at most a

single sample value, and estimation of the between-class variance $\boldsymbol{t}^2$ is not feasible. The

direct regression version of the model (14) with $\boldsymbol{t}^2 = 0$ can be applied in this setting.

Robustness can still be achieved by positing regression models that make weak

parametric assumptions (Breidt and Opsomer 2000; Zheng and Little 2002a). In

particular, Zheng and Little (2002a, 2002b) consider a penalized spline approach based

on the model

24
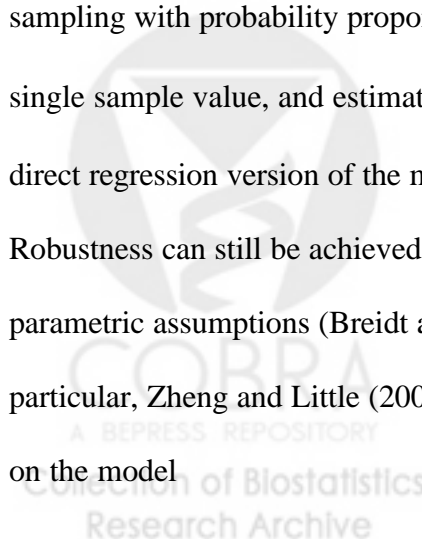
$$y_i = f(\boldsymbol{p}_i, \boldsymbol{b}) + \boldsymbol{e}_i, \ \boldsymbol{e}_i \sim \text{iid} \ G(0, \boldsymbol{p}_i^{2k} \boldsymbol{s}^2), \tag{15}$$

where $\boldsymbol{p}_i$ is the selection probability for unit i, the exponent k (usually taking values 0,1/2 or 1) models error heteroskedasticity, and the function f is a p-spline written as a linear combination of truncated polynomials:

$$\hat{f}(\boldsymbol{p}_i, \boldsymbol{b}) = \boldsymbol{b}_0 + \sum_{j=1}^{p} \boldsymbol{b}_j \boldsymbol{p}_i^{j} + \sum_{l=1}^{m} \boldsymbol{b}_{l+p} (\boldsymbol{p}_i - \boldsymbol{k}_l)_+^p, i = 1,...,N,$$
$$\boldsymbol{b}_{l+p} \underset{iid}{\sim} N(0, \boldsymbol{t}^2), l = 1,...,m. \tag{16}$$

where the constants $\boldsymbol{k}_1 < ... < \boldsymbol{k}_m$ are selected fixed knots and $(u)_+^p = u^p I(u \geq 0)$. The effect of treating $\{\boldsymbol{b}_l, \ l = p+1,..., p+m\}$ as normal random effects is to add a penalty term $\sum_{l=p+1}^{p+m} \hat{\boldsymbol{b}}_l^2 / \boldsymbol{t}^2$ to the sum of squares that is minimized in a least squares fit, thus smoothing their estimates towards zero.

The ability of inferences from this weak model to match or improve on the HT and the GR estimator is illustrated in Tables 1 and 2, which summarize a subset of the simulations in Zheng and Little (2002a,b). Five artificial populations are simulated by adding independent errors with variance 0.2 to the following mean functions relating outcome y_i and the inclusion probabilities $\boldsymbol{p}_i$:
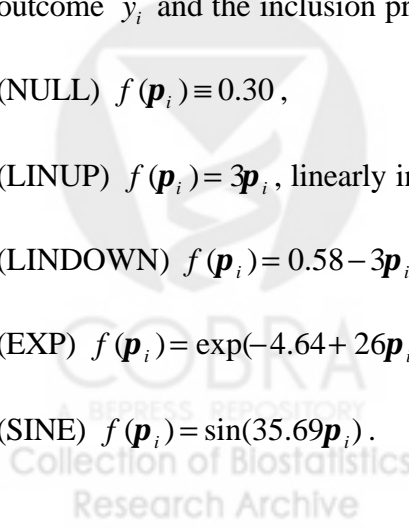
(NULL) $f(\boldsymbol{p}_i) \equiv 0.30$,

(LINUP) $f(\boldsymbol{p}_i) = 3\boldsymbol{p}_i$, linearly increasing function with a zero intercept

(LINDOWN) $f(\boldsymbol{p}_i) = 0.58 - 3\boldsymbol{p}_i$, linearly decreasing function with positive intercept

(EXP) $f(\boldsymbol{p}_i) = \exp(-4.64 + 26\boldsymbol{p}_i)$, an exponentially increasing function

(SINE) $f(\boldsymbol{p}_i) = \sin(35.69\boldsymbol{p}_i)$.

25

A sixth population is generated to yield an "S" shaped function with heteroskedastic errors:

(ESS) $\qquad y_i = 0.6\text{logit}^{-1}(50*p_i - 5 + e_i), e_i \overset{iid}{\sim} N(0,1)$.

Plots of samples from these populations are provided in Figure 1.

Table 1 presents root mean squared error (RMSE) of point estimates from the following methods: HT, the Horvitz-Thompson estimator of the mean; GR, the generalized regression estimator with predictions from a simple linear regression of y_i on p_i, assuming a constant error variance; and P0_15, a p-spline prediction estimator based on (15) and (16) with $k = 0$ and 15 knots. For each of the six mean structures, the RMSE's are based on estimates for 500 systematic samples of size 96 drawn with probability proportional to p_i. Table 1 suggests that P0_15 has smaller empirical RMSE than HT or GR for the populations with nonlinear mean structures (SINE, EXP and ESS). P0_15 has similar RMSE to GR when the mean function is linear (NULL, LINUP and LINDOWN). P0_15 has similar RMSE as HT for the population LINUP, which favors the HT estimator.

Table 2 shows that P0_15, with standard errors computed using the jackknife, yields narrower confidence intervals with coverage properties comparable to that of HT and GR. The only case where P0_15 has poor coverage is the SINE model, and this problem is resolved by increasing the number of knots for the spline. For more details and additional simulation results, See Zheng and Little (2002a, b).

Generalizations of this approach to two-stage sampling are considered in Zheng and Little (2002c). Interestingly, these models lead to improved inferences for two stage

26

samples where the overall probability of selection across the two stage is constant, and the standard estimator is the unweighted mean.

Example 11. Weights in regression revisited. I now return to the question of design and model weights in Example 1, and describe a model that leads to an approximate Bayes estimate that weights by the product of the design and model weights. The basic idea can be conveyed for the simple case of inferences about a mean with no covariates. Assume stratified sampling and the notation in Examples 3 and 8. I first consider a *target* model that is used to define the parameter of interest. This target model assumes the outcomes $\{y_{ji}\}$ in stratum j have a mean that does not depend on stratum, but a non-constant variance, namely

$$p_T(y_{ji} \mid z_{ji} = j, \boldsymbol{q}) \sim G(\boldsymbol{m}, \boldsymbol{s}^2 / u_{ji}), \tag{17}$$

where the notation p_T denotes "target". The target quantity of interest is assumed to be the result of applying this model to the whole population with an uninformative prior, namely the precision-weighted mean:

$$\overline{Y}^{(u)} = \left(\sum_{j=1}^{J} \sum_{i=1}^{N_j} u_{ji} y_{ji} \right) \Big/ \left(\sum_{j=1}^{J} \sum_{i=1}^{N_j} u_{ji} \right). \tag{18}$$

If $u_{ji} = 1$ for all i, j, this is the usual finite population mean, but other choices of $\{u_{ji}\}$ lead to other useful target quantities. For example, if $y_{ji} = x_{ji} / u_{ji}$ then Eq. (17) defines the ratio model, and Eq (18) is the population ratio

$$\left(\sum_{j=1}^{J} \sum_{i=1}^{N_j} x_{ji} \right) \Big/ \left(\sum_{j=1}^{J} \sum_{i=1}^{N_j} u_{ji} \right),$$

27

which is often of substantive interest. A design-based approach might estimate the numerator and denominator of Eq. (18) by design-unbiased estimates, weighting cases in stratum j by the sampling weight $w_j = N_j / n_j$. This yields the estimator

$$\overline{y}^{(w*)} = \left(\sum_{j=1}^{J} \sum_{i \in s_j} w_j u_{ji} y_{ji} \right) \Big/ \left(\sum_{j=1}^{J} \sum_{i \in s_j} w_j u_{ji} \right) = \left(\sum_{j=1}^{J} \sum_{i \in s_j} w_{ji}^* y_{ji} \right) \Big/ \left(\sum_{j=1}^{J} \sum_{i \in s_j} w_{ji}^* \right), \qquad (19)$$

where $w_{ji}^* = w_j u_{ji}$ is the product of the sampling weight and the variance weight. This estimator can also be motivated as an approximate posterior mean under a Bayesian model, as follows. The target model (17) does not reflect the stratified nature of the sample, and as discussed in Example 3, inference for (18) under this model is vulnerable to misspecification, if the means of Y and selection rates vary across the strata. Thus for *inference* about (18), we assume an *inference* model that allows different stratum means, namely

$$
\begin{aligned}
p_I(y_{ji} \mid z_{ji} = j, \boldsymbol{q}) &\sim G(\boldsymbol{m}_j, \boldsymbol{s}_j^{2} / u_{ji}) \\
p\left(\{\boldsymbol{m}_j, \log \boldsymbol{s}_j^{2}\} \right) &= const
\end{aligned}
. \qquad (20)
$$

The possibility of different stratum means is a key feature of the population given the stratified sample design. The inference model yields a posterior predictive distribution for the nonsampled values and hence for the target quantity (18). The resulting inference is not sensitive to violations of the assumptions that the stratum means are constant.

If $\{u_{ji}\}$ are known for all units of the population, a standard Bayesian calculation yields

$$E(\overline{Y}^{(u)} \mid data, \{u_{ji}\}) = \left(\sum_{j=1}^{J} \overline{y}_j^{(u)} \sum_{i=1}^{N_j} u_{ji} \right) \Big/ \left(\sum_{j=1}^{J} \sum_{i=1}^{N_j} u_{ji} \right),$$

28

where $\overline{y}_j^{(u)} = \sum_{i \in s_j} u_{ji} y_{ji} / \sum_{i \in s_j} u_{ji}$ is the precision-weighted mean of the sampled units $i \in s_j$ in stratum j. If $\{u_{ji}\}$ are only known for sampled units of the population, a model is also needed to predict values for nonsampled units. A variety of models for $\{u_{ji}\}$ that involve distinct means in each stratum yield a posterior mean of the total in stratum j of the form

$$E\left(\sum_{i=1}^{N_j} u_{ji} \mid data\right) \doteq w_j \sum_{i \in s_j} u_{ji} ,$$

where $w_j = N_j / n_j$ is the sampling weight for stratum j. Then

$$
\begin{aligned}
E(\overline{Y}^{(u)} \mid data) &= E\left[\left(\sum_{j=1}^{J} \overline{y}_j^{(u)} \sum_{i=1}^{N_j} u_{ji}\right) \Big/ \left(\sum_{j=1}^{J} \sum_{i=1}^{N_j} u_{ji}\right) \Big| data\right] \\
&\doteq \left(\sum_{j=1}^{J} \overline{y}_j^{(u)} E\left\{\sum_{i=1}^{N_j} u_{ji} \mid data\right\}\right) \Big/ \left(\sum_{j=1}^{J} E\left\{\sum_{i=1}^{N_j} u_{ji} \mid data\right\}\right) \\
&= \left(\sum_{j=1}^{J} \overline{y}_j^{(u)} w_j \sum_{i \in s_j} u_{ji}\right) \Big/ \left(\sum_{j=1}^{J} w_j \sum_{i \in s_j} u_{ji}\right) = \sum_{j=1}^{J} \sum_{i \in s_j} w_{ji}^* y_{ji} \Big/ \sum_{j=1}^{J} \sum_{i \in s_j} w_{ji}^* = \overline{y}^{(w*)} ,
\end{aligned}
$$

the estimator (19). The approximation in the second line of this expression results from approximating the posterior expectation of a ratio by a ratio of posterior expectations, which ignores terms of order $O(1/n)$. Hence under this formulation, the model-based approach and design based approach both lead to multiplying the model and design weights together, unifying the two approaches to inference.

An extension of this analysis yields estimates for regression coefficients. Consider more generally the target regression model

$$(Y \mid X, \boldsymbol{b}) \sim G(X\boldsymbol{b}, U^{-1}\boldsymbol{s}^2) , \tag{21}$$

29

where Y consists of the population elements as an $(N \times 1)$ vector, X is an $(N \times p)$ matrix of covariates, and U is a $(N \times N)$ diagonal matrix with the value $\{u_{ji}\}$ on the diagonal. The target quantities are the precision-weighted least squares estimates:

$$B^{(u)} = (X^T U X)^{-1} X^T U Y. \tag{22}$$

For inference about (22), we assume an inference model that allows different stratum regression coefficients, namely

$$\begin{aligned}
(Y_j \mid X_j, \boldsymbol{b}_j, \boldsymbol{q}) &\sim G(X_j \boldsymbol{b}_j, U_j^{-1} \boldsymbol{s}_j^2) \\
p\left(\{\boldsymbol{b}_j, \log \boldsymbol{s}_j^2\}\right) &= const
\end{aligned} \tag{23}$$

where Y_j, X_j are the components of Y and X in stratum j, with dimension $(N_j \times 1)$ and $(N_j \times p)$ respectively. An approximation to the posterior mean of $B^{(u)}$ under (23) is obtained by writing (22) as a function of sums

$$B^{(u)} = g(T_1, \ldots, T_L),$$

where $\{T_\ell = \sum_{j=1}^{J} \sum_{i=1}^{N_j} u_{ji} h_{\ell ji}, \ell = 1, \ldots, L\}$, for difference choices of $\{h_{\ell ji}\}$ represent the set of sums, sums of squares, and sums of cross products of the covariates and outcome. Then
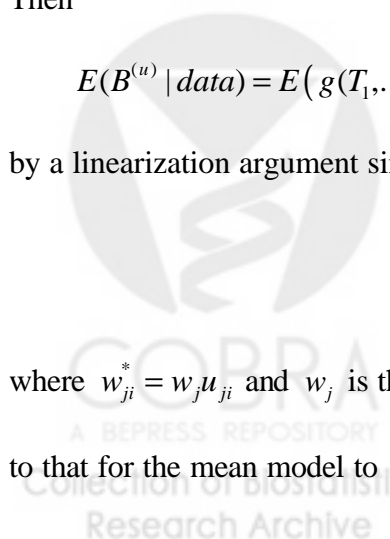
$$E(B^{(u)} \mid data) = E\left(g(T_1, \ldots, T_L) \mid data\right) \simeq g\left(E(T_1 \mid data), \ldots, E(T_L \mid data)\right) + O(1/n),$$

by a linearization argument similar to that used for design-based inference. Also,

$$E(T_\ell \mid data) \simeq \sum_{j=1}^{J} \sum_{i \in s_j} w_{ji}^* h_{\ell ji},$$

where $w_{ji}^* = w_j u_{ji}$ and w_j is the sampling rate in stratum j, applying an argument similar to that for the mean model to $\{h_{\ell ji}\}$. Hence:

$$E(B^{(u)} \mid data) \simeq (X_s{}^T W_s{}^* X_s)^{-1} X_s{}^T W_s{}^* Y_s, \qquad (24)$$

where the subscript s denotes sample quantities. This analysis generalizes the results in Little (1991), who considers the constant variance case where $u_{ji} = 1$ for all i, j.

Can sampling weights be ignored when interest lies in "analytic" inference for the parameters $\boldsymbol{b}$ of the target model (21), rather than in the finite population quantity (22)? I would say no, Eq. (24) should still be used to estimate $\boldsymbol{b}$. The inference differs only in the omission of finite population corrections, which follows directly from the application of Bayes theorem. My reason is that the finite population is assumed a random sample from the superpopulation under the superpopulation model, so $\boldsymbol{b}$ differs from the finite population quantity $B^{(u)}$ by a (small) quantity of order $O(1/N)$. Since ignoring the sampling weights yields a poor estimate of $B^{(u)}$, it also yields a poor estimate of $\boldsymbol{b}$.

A design-based statistician might ask what is gained by modeling if the "robust" model (23) merely recovers the design-based estimator. My answer is that as in Examples 3 and 9, the Bayesian paradigm allows for better small-sample inferences, by propagating error in estimating the variances, and by allowing the possibility of shrinkage of the weights by mixed models such as (14). An examination of these gains is a topic for future research.

6. Conclusion

In this article I have reviewed some aspects of the debate between design-based on model-based inference for sample surveys. An interesting question is the extent to which we are now in (or entering) an "age of reconciliation" between design-based and model-based approaches (Smith 1994) after the "hundred years war" (Kish 1995). In practical terms this may be true, particularly in large-sample settings. Modelers are more

31

focused on the impact of the design, and design-based survey samplers are paying more attention to models, as in the methods in Section 4. Models that take into account design features will tend to yield similar results to design-based methods that are sensitive to the implied model for the population values.

On the other hand, the theoretical underpinnings of the two approaches remain strikingly different, and I think theory matters. My own position is that the Bayesian paradigm is flexible enough to provide practical and useful inferences for data collected by sample surveys, as with data collected by other selection mechanisms. However, models need to properly reflect features of the sample design such as weighting, stratification and clustering, or inferences are likely to be distorted.

In this article I focused mainly on point estimation, and have not discussed estimation of precision. In principle I prefer estimates of precision to be based on the Bayesian posterior distribution for a carefully specified model, but other methods of precision estimation that trade efficiency for robustness, such as replication methods and the "sandwich" estimator, have some appeal in the production survey setting, where sample sizes are large and detailed model assessment is not practical. Emphasis should be on the properties of inferences themselves, such as confidence intervals or P-values, rather than on intermediate quantities such as variance estimates.

I conclude by addressing two other criticisms of the model-based approach by advocates of design-based inference. The first is that modelers don't believe in random sampling, since the sampling distribution is not the basis for inference. As noted in Section 2, a model-based approach that ignores the sampling mechanism is not valid unless the sampling distribution does not depend on the survey outcomes. Otherwise, the

32

sampling mechanism needs to be modeled, and appropriate modeling in such cases is at best difficult. Probability sampling is amply justified within the modeling paradigm by the need for robustness to model misspecification

Another criticism of the model-based approach is that it is impractical for large-scale survey organizations: the work in developing good models, and the computational complexity of fitting them, is not suited to the demands of "production-oriented" survey analysis. However, attention to models is needed in model-assisted approaches, even when the basis for inference is the sample design. Also, computational power has expanded dramatically since the days of early model versus randomization debates, and much can be accomplished using software for mixed models in the major statistical packages (SAS 1992; Pinheiro and Bates 2000) or Bayesian software based on MCMC methods such as BUGS. (Spiegelhalter, Thomas, and Best 1999). Bayesian software targeted at complex survey problems would increase the utility of this approach for practitioners. Also, guidance on "off-the-shelf" models for routine application to standard sample designs would be useful, although no statistical procedure, design or model-based, should be applied blindly without any attention to diagnostics of fit to the data.

33
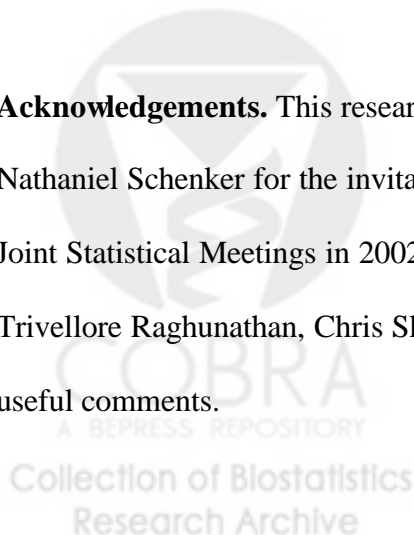
References

Basu, D. (1971), "An essay on the logical foundations of survey sampling, Part 1," in *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston, pp. 203-242.

Binder, D. A. (1982), "Non-parametric Bayesian Models for Samples from Finite Populations," *Journal of the Royal Statistical Society* 44, 3, 388-393.

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Sample Surveys. *International Statistical Review* 51, 279-92.

Breidt, F.J. and Opsomer, J.D. (2000), "Local Polynomial Regression Estimators in Survey Sampling," *Annals of Statistics* 28, 1026-53.

Brewer, K. R. W. (1979), "A Class of Robust Sampling Designs for Large-Scale Surveys," *Journal of the American Statistical Association* 74, 911-915

Brewer, K.R.W. and Mellor, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys," *Australian Journal of Statistics* 15, 145-152.

Cassel, C-M, Särndal, C-E. and Wretman, J.H. (1977), *Foundations of Inference in Survey Sampling*, New York: Wiley.

Chambers, R.L. (2003), "Introduction to Part A," in *Analysis of Survey Data*, R.L. Chambers and C.J. Skinner, eds., New York: Wiley, pp.13-27.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd Edition, New York: Wiley.

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, London: Chapman Hall.

De Finetti, B. (1990), *Theory of Probability. A Critical Introductory Treatment* (Vols 1 and 2), New York: Wiley.

34
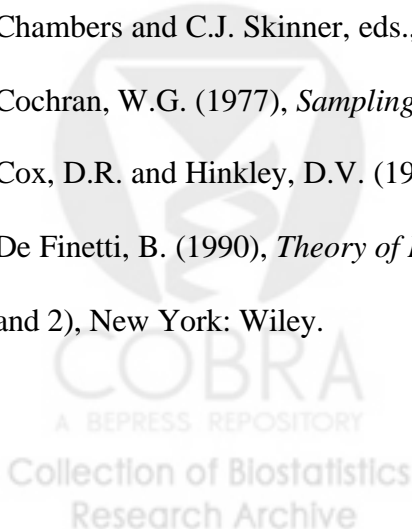
Dumouchel, W.H. and Duncan, G.J. (1983), "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples," *Journal of the American Statistical Association* 78, 535-543.

Elliott, M. R. and Little, R.J.A. (2000). Model-Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics* 16, No. 3, 191-209.

Ericson, W.A. (1969), "Subjective Bayesian Models in Sampling Finite Populations," *Journal of the Royal Statistical Society*, B 31, 195-234.

Ericson, W. A. (1988), "Bayesian Inference in Finite Populations," in *Handbook of Statistics* 6, Amsterdam: North-Holland, pp. 213-246,

Firth, D. and Bennett, K.E. (1998), "Robust Models in Probability Sampling," *Journal of the Royal Statistical Society*, B 60, 3-21.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs' Sampling," *Journal of the American Statistical Association*, 85, 972-985.

Ghosh, M. and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*. London: Chapman & Hall.

Godambe, V.P. (1955), "A Unified Theory of Sampling from Finite Populations," *Journal of the Royal Statistical Society*, B 17, 269-278.

Godambe, V.P. and Thompson, M.E. (1986), "Parameters of Superpopulations and Survey Population: their Relationship and Estimation. *International Statistical Review* 54, 37-59.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sampling Survey Methods and Theory*, Vols. I and II, New York: Wiley.

35

Hansen, M.H. and Hurwitz, W.N. (1943), "On the Theory of Sampling from Finite Populations," *Annals of Mathematical Statistics* 14, 333-362.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," *Journal of the American Statistical Association* 78, 776-793 (with discussion).

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

Holt, D., Smith, T.M.F., and Winter, P.D. (1980), "Regression Analysis of Data from Complex Surveys," *Journal of the Royal Statistical Society*, A 143, 474-87.

Horvitz, D.G., and Thompson, D.J. (1952), "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association* 47, 663-685.

Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model", *Journal of the American Statistical Association* 77, 89-96.

Kalton, G. (2002), "Models in the Practice of Survey Sampling (Revisited)," *Journal of Official Statistics* 18, 129-154.

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

Kish, L. (1995). "The Hundred Years' Wars of Survey Sampling," *Statistics in Transition*, 2, 813-830. Reproduced as Chapter 1 of G. Kalton and S. Heeringa, (2003, eds.) *Leslie Kish: Selected Papers*, New York: Wiley.

Kish, L. and Frankel, M.R. (1974), "Inferences from Complex Samples (with discussion), *Journal of the Royal Statistical Society* B 36, 1-37.

36

Konijn, H.S. ((1962), "Regression Analysis in Sample Surveys," *Journal of the American Statistical Association* 57, 590-606.

Korn, E.L. and Graubard, B.I. (1999), *Analysis of Health Surveys*, New York: Wiley.

Lazzeroni, L.C., and Little, R.J.A. (1998), "Random-Effects Models for Smoothing Post-Stratification Weights," *Journal of Official Statistics* 14, 61-78.

Little, R.J.A. (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association* 77, 237-250.

Little, R.J.A. (1991), "Inference with Survey Weights," *Journal of Official Statistics* 7, 405-424.

Little, R.J.A. (1993), "Post-Stratification: a Modeler's Perspective," *Journal of the American Statistical Association* 88, 1001-1012.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data,* 2nd edition, New York: Wiley.

Mahalanobis, P.C. (1943), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society* 109, 325-378.

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*, A 97, 558-606.

Pfeffermann, D. (1993), "The Role of Sampling Weights when Modeling Survey Data," *International Statistical Review* 61, 317-337.

Pfeffermann, D., Skinner, C.J., Homes, D.J., Goldstein, H. and Rasbach, J. (1998). "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society*, B 60, 23-40.
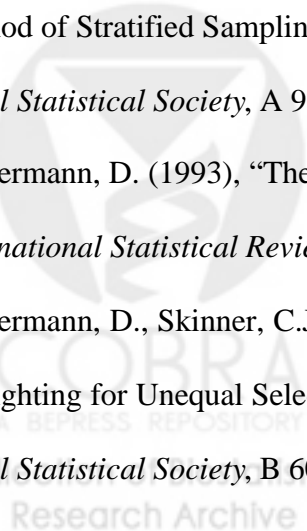
Pfeffermann, D. and Holmes, D.J. (1985), "Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data," *Journal of the Royal Statistical Society*, A 148, 268-278.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plus*, New York: Springer.

Rizzo, L. (1992), "Conditionally Consistent Estimators Using Only Probabilities of Selection in Complex Sample Surveys," *Journal of the American Statistical Association*, 87, 1166-1173.

Royall, R. M. (1970), "On Finite Population Sampling Under Certain Linear Regression Models," *Biometrika* 57, 377-387.

Rubin, D.B. (1976), "Inference and Missing Data," *Biometrika* 53, 581-592.

Rubin, D.B. (1983), Comment on "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," by M.H. Hansen, W.G. Madow, and B.J. Tepping, *Journal of the American Statistical Association* 78, 803-805.

Rubin, D.B. (1985), "The Case of Propensity Scores in Applied Bayesian Inference", in *Bayesian Statistics*, eds. I.M. Bernado, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Amsterdam: Elsevier, pp. 463-472.

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

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

SAS (1992), "The Mixed Procedure," in *SAS/STAT Software: Changes and Enhancements, Release 6.07,* Technical Report P-229, SAS Institute, Inc., Cary, NC.

Scott, A.J. (1977), "Large-Sample Posterior Distributions for Finite Populations," *Annals of Mathematical Statistics*, 42, 1113-1117.

Scott, A.J. and Smith, T.M.F. (1969), "Estimation in Multistage Samples," *Journal of the American Statistical Association* 64, 830-840.

Smith, T.M.F. (1976), "The foundations of Survey Sampling: a Review," *Journal of the Royal Statistical Society*, A 139, 183-204 (with discussion).

Smith, T.M.F. (1988), "To Weight or not to Weight, that is the Question," in *Bayesian Statistics 3*, J.M. Bernado, M.H. DeGroot and D.V. Lindley, eds., Oxford, U.K.: Oxford University Press, pp. 437-451.

Smith, T.M.F. (1994), "Sample Surveys 1975-1990; an Age of Reconciliation?" *International Statistical Review* 62, 5-34 (with discussion).

Spiegelhalter, D.J., Thomas, A. and Best, N.J. (1999), *WinBUGS Version 1.2 User Manual*, MRC Biostatistics Unit, Cambridge, UK.

Thompson, M.E. (1988), ``Superpopulation Models", *Encyclopedia of Statistical Sciences (Vol. 1)* 9, 93-99 .

Valliant, R. , Dorfman, A.H. , and Royall, R. M. (2000), *Finite Population Sampling and Inference: a Prediction Approach*, New York: Wiley.

Zheng, H. and Little, R.J.A. (2002a), Penalized spline model-based estimation of the finite population total from probability proportional to size samples. To appear in *Journal of Official Statistics.*

Zheng, H. and Little, R.J.A. (2002b), Inference for the Population Total from Probability-Proportional-To-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. Submitted for publication.
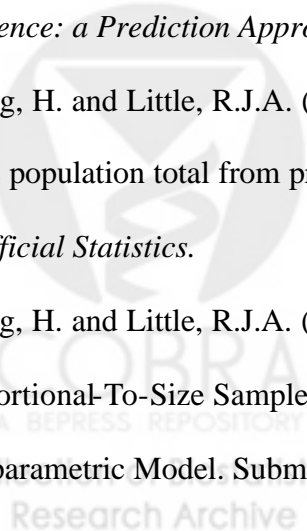
Zheng, H. and Little, R.J.A. (2002c), Penalized Spline Nonparametric Mixed Models for Inference About a Finite Population Mean from Two-Stage Samples. Submitted for publication.

Figure 1. Six simulated populations (N=300) X-axis: pi(i); Y-axis: y(i) with normal
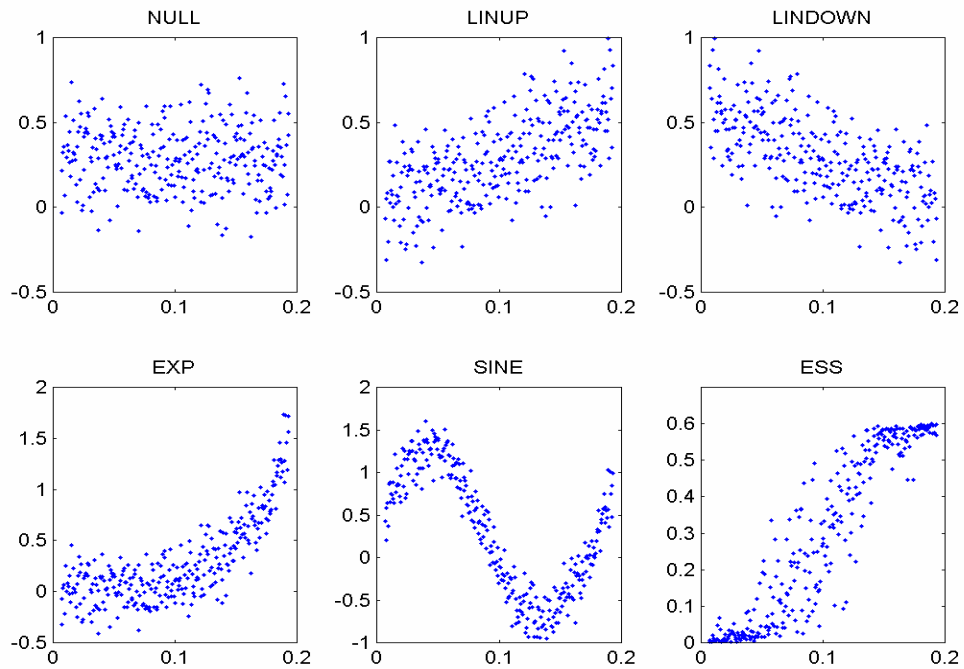
errors

Table 1. RMSE of three point estimators: P0_15, HT and GR
N=1000,n=96

	HT	GR	P0_15
NULL	35	24	22
LINUP	27	34	26
LINDOWN	63	35	27
SINE	113	95	45
EXP	35	54	27
ESS	11	30	10

Table 2. Average Width (AW) and Noncoverage rate (NC) of 95% C.I.s over 1000 samples (target 50 +/- 20). Comparisons of HT = Horvitz Thompson with random groups variance estimate, GR = Generalized Regression with Yates-Grundy variance estimate, P0_15 = P-spline with Jackknife variance estimate. N = 1000, n = 100.

	HT		GR		P0_15	
	AW	NC	AW	NC	AW	NC
NULL	131	68	88	80	89	28
LINUP	109	42	123	64	98	48
LINDOWN	230	82	124	82	94	62
SINE	446	60	340	74	145	86
EXP	135	42	193	96	105	54
ESS	48	14	109	84	37	66