Objective Stepwise Bayes Weights in Survey Sampling

Jeremy Strief

Principal Statistician

Medtronic Energy and Component Center

Brooklyn Center, MN 55430

jstrief@gmail.com

Glen Meeden*

School of Statistics

University of Minnesota

Minneapolis, MN 55455

glen@stat.umn.edu

Submitted October 2011
Revised August 2012

*Research supported in part by NSF Grant DMS 0406169

SUMMARY

Although weights are widely used in survey sampling their ultimate justification from the design perspective is often problematical. Here we will argue for a stepwise Bayes justification for weights that does not depend explicitly on the sampling design. This approach will make use of the standard kind of information present in auxiliary variables however it will not assume a model relating the auxiliary variables to the characteristic of interest. The resulting weight for a unit in the sample can be given the usual interpretation as the number of units in the population which it represents.

AMS 1991 subject classifications Primary 62D05; secondary 62F15.

Key Words and phrases: sample survey, weights, Bayesian inference

1 Introduction

Weights play an important role in the design based approach to survey sampling. In theory the weight assigned to an observed unit in a sample is the reciprocal of its selection probability and is interpreted as the number of units in the population which it represents. In practice, after a sample has been observed, the weights are often adjusted to make the sample better represent the population. These adjustments can be made to take into account population information not included in the design and for observations missing from the sample. Although such modifications of the design based weights are undoubtedly useful in some cases their ultimate theoretical justification is not so clear. Part of the confusion, we believe, comes from arguing unconditionally before the sample is taken, e.g. the Horvitz-Thompson estimator is unbiased averaged over all possible samples, and then conditionally after the sample is in hand, by adjusting the designed based weights of the observed units in the sample. In particular, an overemphasis on the sampling design at the second or conditional stage can needlessly complicated matters. After the sample has been observed, we believe a better approach is to formally ignore the sampling design but use all the available information, including that embedded in the design, to find a sensible set of weights. In this way of thinking a weight assigned to a unit can still be interpreted as the number of units in the population that it represents but it is no longer derived as an adjustment of its selection probability. How can this be done?

In the Bayesian approach information about the population is incorporated into a prior distribution. In theory, the prior can then be used to purposely select an optimal sample; however this is almost never done. After the sample is observed inferences are based on the posterior distribution of the unobserved units in the population given the values of the observed units in the sample. In most situations the posterior does not depend on how the sample was selected and hence the design plays no role at the inference stage. Bayes methods have been little used in practice because it is difficult to find prior distributions which reflect the common kinds of available prior information.

Many of the standard estimators can be given a stepwise Bayesian interpretation (Ghosh and Meeden, 1997). In this approach, given any sample, inference is still based on a posterior distribution but the collection (for all possible samples) of the posteriors does not arise from a single prior but from a whole family of prior distributions. In the situation where one believes that the observed units are roughly exchangeable with the unobserved units the appropriate stepwise Bayes posterior distribution is the Polya posterior.

When prior information about population means and quantiles of auxiliary variables is available Lazar et al. (2008) argued that the constrained Polya posterior, a generalization of the Polya posterior, is a sensible way to incorporate such prior information. Here we will show how the constrained Polya posterior can be used to define weights for the units in the sample. Although the resulting weights depend on the auxiliary variables they do not make explicit use of the sampling design.

In sections 2 and 3 we review the Polya posterior and the constrained Polya posterior. The two main ideas of the paper are given in the next two sections. In section 4 we show how the constrained Polya posterior can be used to attached a weight to each unit in the sample and in such a way that these weights do not depend directly on the sampling design. In section 5 we introduce the weighted Dirichlet posterior as a companion to the constrained Polya posterior. It allows one to use the weights defined by the constrained Polya posterior to make inferences about population parameters through straight forward simulation. In section 6 we compare the constrained Polya posterior weights to those used in the Horvitz-Thompson estimator. In section 7 we consider several examples to see how the resulting weights preform in practice and show how the weighted Dirichlet posterior can be use to get an estimate of variance for an estimator without extensive computing. Section 8 contains some concluding remarks.

At first reading it will seem to some that the methods proposed here are very Bayesian because all of our inferences are based on "posterior" distributions. But as mentioned above, technically, our "posterior" distributions are not Bayes but stepwise Bayes. This means that operationally one can think of our posterior as being constructed after the sample has been observed. These constructed "posteriors" do not depend on subjective prior information or the sampling design but just use the observed sample values and objective and public information about the auxiliary variables. As we shall see this allows one to construct estimators of population parameters which are approximately unbiased under a variety of designs and have good frequentist properties. There are two important limitations of our work however. The first is that it only is applicable to single stage designs and the second is that it cannot correct for selection bias..

2 The Polya posterior

Let s be the set of labels of a sample of size n from a population of size N. For convenience we assume the members of s are $1, 2, \ldots, n$ and we also suppose that n/N is very small. Let $y = (y_1, y_2, \ldots, y_N)$ be the characteristic of interest and y_s be the observed sample values.

The Polya posterior is based upon Polya sampling from an urn. Polya sampling works as follows: suppose that the values from n observed or seen units are marked on n balls and placed in urn 1. The remaining unseen $N - n$ units of the population are represented by $N - n$ unmarked balls placed in urn 2. One ball from each urn is drawn with equal probability, and the ball from urn 2 is assigned the value of the ball from urn 1. Both balls are then returned to urn 1. Thus at the second stage of

Polya sampling, urn 1 has $n+1$ balls and urn 2 has $N-n-1$ balls. This procedure is repeated until urn 2 is empty, at which point the N balls in urn 1 constitute one complete simulated copy of the population. Any finite population quantity—means, totals, quantiles, regression coefficients—may now be calculated from the complete copy. For the population quantity of interest we may simulate K such complete copies and in each case calculate its value. The mean of these simulated values is the point estimate and an approximate 95% Bayesian credible interval is given by the 2.5% and 97.5% quantiles of the values.

One can check that under the Polya posterior the posterior expectation of the population mean is just the sample mean and the posterior variance is just $(n-1)/(n+1)$ times the usual design based variance of the sample mean under simple random sampling without replacement. The Polya posterior has a decision theoretic justification based on its stepwise Bayes nature. Using this fact many standard estimators can be shown to be admissible. Details can be found in Ghosh and Meeden (1997). The Polya posterior is the Bayesian bootstrap of Rubin (1981) applied to finite population sampling. Lo (1988) also discusses the Bayesian bootstrap in finite population sampling. Some early related work can be found in Hartley and Rao (1968) and Binder (1982).

For the sample unit i let p_i denote the proportion of units in a full, simulated copy of the population which have the value y_i. Ghosh and Meeden (1997) showed that under the Polya posterior $E(p_i) = 1/n$. If we let

$$w_i = NE(p_i) = N/n$$

then w_i can be interpreted as the weight attached to unit i since it equals the average number of units in the population represented by unit i, under the Polya posterior. Recall that under simple random sampling without replacement n/N is the inclusion probability for each unit. Hence in this case the usual frequentist weight, which is the reciprocal of the inclusion probability, and Polya posterior weight defined above agree.

So in situations of limited prior information the Polya posterior yields weights identical to frequentist weights derived from the design of simple random sampling without replacement. The Polya posterior justification for these weights does not depend explicitly on the design and would be appropriate anytime the sampler believes the observed and unobserved units in the population are roughly exchangeable.

We next address the issue of the relationship of the Polya posterior with usual bootstrap methods in finite population sampling. Both approaches are based on an assumption of exchangeability. Gross (1980) introduced the basic idea for the bootstrap. Assume simple random sampling without replacement and suppose it is the case that $N/n = m$ is an integer. Given a sample we create a good guess for the population by combining m replicates of the sample. By taking repeated random samples of size n from this created population we can study the behavior of an estimator of interest. Booth et al. (1994) studied the asymptotic properties of such estimators. Hu et al. (1997) is an example where the sample was used to construct an artificial population and then repeated samples were drawn from the constructed population to construct an estimate of the variance of their estimator and to construct confidence intervals.

Note this is in contrast to the Polya posterior which considers the sample fixed and repeatedly generates complete versions of the population.

3 The constrained Polya posterior

We begin by recalling a well known approximation to the Polya posterior. If n/N is small then under the Polya posterior, $p = (p_1, \ldots, p_n)$ has approximately a Dirichlet distribution with a parameter vector of all ones, i.e., it is uniform on the $n-1$ dimensional simplex, where $\sum_{j=1}^{n} p_j = 1$. It is usually more efficient to generate complete copies of the population using this approximation than the urn model described in the previous section. In addition this approximation will be useful when we consider the constrained Polya posterior, a generalization of the Polya posterior. which arises when prior information about auxiliary variables are available to the sampler.

In many problems, in addition to the variable of interest, y, the sampler has in hand auxiliary variables for which prior information is available. A very common case is when the population mean of an auxiliary variable is known. More generally, we will assume that prior information about the population can be expressed by a set of linear equality and inequality constraints on a collection of auxiliary variables.

We assume that in addition to the characteristic of interest y there is a set of auxiliary variables $x^1, x^2, \ldots, x^m$. For unit i let

$$(y_i, x_i) = (y_i, x_i^1, x_i^2, \ldots, x_i^m)$$

be the vector of values for y and the auxiliary variables. We suppose that for any unit in the sample this vector of values is observed. We assume the prior information about the population can be expressed through a set of linear equality and inequality constraints on the population values of the auxiliary variables. For the set of possible values for a given auxiliary variable the coefficients defining a constraint will correspond to the proportions of units in the population taking on these values. We now illustrate this more precisely by explaining how we translate this prior information about the population to the observed sample values. Given a sample this will allow us to construct simulated copies of the population consistent with the prior information.

Given a sample s, for $i = 1, 2, \ldots, n$, let (y_i, x_i) be the observed values which, for simplicity, we assume are distinct. Let p_i be the proportion of units which are assigned the value (y_i, x_i) in a simulated complete copy of the population. Any linear constraint on the population value of an auxiliary variable translates in an obvious way to a linear constraint on these observed values. For example, if the population mean of x^1 is known to be less than or equal to some value, say b_1, then for the simulated population this translates to the constraint

$$\sum_{i=1}^{n} p_i x_i^1 \leq b_1$$

If the population median of x^2 is known to be equal to b_2 then for the simulated population this becomes the constraint

$$\sum_{i=1}^{n} p_i u_i = 0.5$$

where $u_i = 1$ if $x_i^2 \leq b_2$ and it is zero otherwise. Hence, given a collection of population constraints based on prior information and a sample we will be able to represent the corresponding constraints on a simulated value of p by two systems of equations

$$A_{1,s}\, p = b_1 \tag{1}$$

$$A_{2,s}\, p \leq b_2 \tag{2}$$

where $A_{1,s}$ and $A_{2,s}$ are $m_1 \times n$ and $m_2 \times n$ matrices and b_1 and b_2 are vectors of the appropriate dimensions.

Let $\mathcal{P}$ denote the subset of the n dimensional simplex which is defined by equations 1 and 2. We assume the sample is such that $\mathcal{P}$ is non-empty and hence it is a non-full dimensional polytope. In this case the appropriate approximate version of the Polya posterior should just be the uniform distribution over $\mathcal{P}$. We call this distribution the constrained Polya posterior (CPP). If one could generate independent observations from the CPP then one could find approximately the posterior expectation of population parameters of interest and find approximate 0.95 stepwise Bayes credible intervals. Unfortunately we do not know how to do this. Instead, one can use Markov chain Monte Carlo (MCMC) methods to find such estimates approximately. This can done in R (R Development Core Team, 2005). and using the R package *polypost* which is available in CRAN. More details on the CPP and simulating from it are available in Lazar et al. (2008).

4 Constrained Polya posterior weights

A possible criticism of the Polya posterior and the CPP is that any simulated full copy of the population will only contain values of the characteristic that appeared in the sample. But it is exactly this property that will allow us to attach weights to the members of the sample.

We assume that we have a fixed sample for which the subset of the simplex defined by equations 1 and 2 is nonempty. For $j = 1, \ldots, n$ let

$$w_j = NE(p_j) = N\mu_j \tag{3}$$

where the expectation is taken with respect to the CPP. Note that the sum of the elements of $w = (w_1, \ldots, w_n)$ is the population size N and w_j can be thought of as the weight associated with the jth member of the sample. These weights depend only on the observed values of the auxiliary variables and the known population constraints. Hence this is a stepwise Bayes method of attaching weights to the units in the sample which incorporates the prior information present in the auxiliary variables and does not depend explicitly on the sampling design.

We are assuming here that the population size N is know which may not always be the case. In such situations one could replace N in the above equation by an estimate. If the estimate is a good one then the resulting inferences for a population total should be satisfactory. When estimating a population mean the results would be much less sensitive to how close the estimate is to the true population size.

Much survey data which are used by social science researchers comes with weights attached to individual units. In such cases the CPP weights could be attached in the same way and the user would not need to use MCMC methods to calculate the weights. We will use the weights to define the Weighted Dirichlet posterior that can be used to find point and interval estimates of population quantities of interest at a relative modest computational cost. In the rest of the paper we will give examples to show that these weights can be used to generate inferential procedures with good frequentist properties.

But before proceeding we make a simple observation. Suppose we have in hand the sample along with a set of weights. If N is large, then we can construct a population where the proportion of units in the population of type (y_i, x_i) is w_i/N for $i = 1, \ldots, n$. Given the sample and the set of weights, we can think of this constructed population as the best guess for the unknown population. Then

$$\bar{y}_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} y_i \quad \text{and} \quad \sigma^2_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} (y_i - \bar{y}_{bw})^2 \tag{4}$$

are the mean and variance of this constructed population.

5 The weighted Dirichlet posterior

It is often the case that weights are attached to data in public use files. These weights are then used by researchers to make point and interval estimates of population parameters. We shall see that the stepwise Bayes weights introduced here can often be used in standard frequentist formulas to estimate parameters of interest just as the usual weights are. We will use our weights to define the Weighted Dirichlet posterior (WDP) and show that it gives an alternative way to compute point and interval estimates for a variety of population quantities.

Let the w_j's be a set of weights defined by equation 3 with $\mu_j = w_j/N$. Consider the Dirichlet distribution over the simplex defined by the vector $n\mu = (n\mu_1, \ldots, n\mu_n)$ as an alternative posterior distribution for $p = (p_1, \ldots, p_n)$ when using the observed sample to generate complete simulated copies of the population. We will call this posterior the weighted Dirichlet posterior (WDP). Note the WDP is a looser version of the CPP. Under the CPP every complete copy of the population will satisfy the constraints; however, under the WDP, only the average of all the simulated populations will satisfy the constraints. It is easy to see that under the WDP

$$E\Big(\sum_{i=1}^{n} p_i y_i\Big) = \sum_{i=1}^{n} \mu_i y_i = \bar{y}_{bw} \tag{5}$$

and

$$\begin{aligned} V\Big(\sum_{i=1}^{n} p_i y_i\Big) &= \sum_{i=1}^{n} y_i^2 V(p_i) + \sum\sum_{i<j} y_i y_j Cov(p_i, p_j) \\ &= \sum_{i=1}^{n} \frac{n\mu_i(n-n\mu_i)y_i^2}{n^2(n+1)} - 2\sum\sum_{i<j} \frac{n\mu_i n\mu_j y_i y_j}{n^2(n+1)} \\ &= \frac{1}{n+1}\Big(\sum_{i=1}^{n} \mu_i(1-\mu_i)y_i^2 + 2\sum\sum_{i<j} \mu_i n\mu_j y_i y_j\Big) \\ &= \frac{1}{n+1}\Big(\sum_{i=1}^{n} \mu_i y_i^2 - \sum_{i=1}^{n}\sum_{i=1}^{n} \mu_i\mu_j y_i y_j\Big) \\ &= \frac{1}{n+1}\sigma_{bw}^2 \end{aligned} \tag{6}$$

where $\bar{y}_{bw}$ and σ_{bw}^2 were defined in equation 4.

From this we see that when estimating the population mean, simulating from the WDP is equivalent to using the sample and their weights to construct the best guess for the population. In particular, when the weights are all equal the WDP is just the Polya posterior.

There are two main reasons for introducing the WDP. The first is that as the number of constraints used increases the approximate 0.95 credible intervals based on the CPP become too short and contain the true parameter value less than 95% of the time. This happens because with a large number of constraints the CPP does not allow enough variability in the simulated complete copies of the population which it generates. The second reason is that simulating from the WDP is much easier that simulating from the CPP. Now it would be possible to simulated from the constrained WDP in such a way that all the constraints would be satisfied but this involves as much effort as simulating from the CPP. Moreover, we believe that this would yield approximate 0.95 credible intervals which have poor frequentist coverage properties because they are too short.

Now suppose our set of weights is the reciprocals of the inclusion probabilities from the sampling design. Let $W = \sum_{i=1}^n w_i$. For most samples this value will not be equal to N but often is is quite close. Again we can construct our best guess for the population based on the weights. The mean and variance of this population will be

$$\bar{y}_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i \quad \text{and} \quad \sigma^2_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} (y_i - \bar{y}_{dw})^2 \tag{7}$$

If we use $\bar{y}_{dw}$ as an estimate of the unknown population mean then an unbiased estimate of its variance depends on the joint inclusion probabilities of the units in the sample. Since these are often difficult to obtain, what has been recommended in practice (Särndal et al., 1992) is to assume the sampling was done with replacement even when that is not the case. Then the resulting approximate estimate of variance for $\bar{y}_{dw}$ is

$$\begin{aligned} \hat{V}_d(\bar{y}_{dw}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \big(n \frac{w_i}{W} y_i - \bar{y}_{dw}\big)^2 \\ &= \frac{\sigma^2_{dw} + \gamma_{dw}}{n-1} \end{aligned} \tag{8}$$

where the second line follows from some simple algebra and where

$$\gamma_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i^2 \big(n \frac{w_i}{W} - 1\big) \tag{9}$$

Note that when the design is simple random sampling with or without replacement and $N = nk$ then $\gamma_{dw} = 0$ In this case, this estimate of variance is essentially equivalent to the variance in equation 6.

In situations where the Horvitz-Thompson estimator makes sense, calculations have shown that γ_{dw} tends to be negative. This suggests that in such situations intervals based on the WDP will tend to be conservative. However calculations also show that γ_{dw} term tends to be positive in situations where the Horvitz-Thompson estimator is not appropriate. We will see in such cases that the usual approximation can work poorly and intervals based on the WDP can have better frequentist properties.

6 Weights and Horvitz-Thompson

The usual definition of the weight assigned to a unit in the sample is the inverse of its inclusion probability. One is encouraged to think of a unit's weight as being the number of units in the population which it represents. The resulting estimator of the population total is the Horvitz-Thompson (HT) estimator and is design unbiased. As we have already noted the unbiased estimate of its variance depends on the joint selection probabilities of the all the pairs of units appearing in the sample. Since in practice this can be impossible to compute the approximation in equation 8 is often used.

The HT estimator works best when y_i is approximately proportional to its selection probability. To compare its behavior to the WDP method we conducted a small simulation experiment. We constructed the variable x by drawing a random sample of 2,000 from from a gamma distribution with shape parameter 5 and scale parameter 1 and adding 20 to each value. To generate y we let the conditional distribution of y_i given x_i be a normal distribution with mean $5x_i$ and standard deviation 20. The correlation of the resulting population was 0.49. We denoted this population by A. We created a second population, B, by using the same vector of x values but adding 400 to each y_i value. Our sampling plan used x to do sampling proportional to size, i.e. $pps(x)$. We used the R package *sampling* so that the inclusion probabilities were exact. Under this design we expect that the HT estimator would work well for population A but perform less well for population B. We also considered a third estimator, NHT, which is just the weights of the HT estimator rescaled so that they sum to the population size. We generated 500 samples of size 50. The results are giving in Table 8

Although not shown in the table both the HT and WDP estimators are unbiased for both populations. As expected the HT estimator is the best for population A although its performance falls off dramatically for population B. On the other hand the WDP performance for both populations is exactly the same. As a point estimator the NHT does much better than the HT estimator for population B but not as well for population A. Overall the WDP is clearly performs the best. What is an explanation for these differences?

In population A, $y_i \propto x_i$ and calculations show that γ_{dw} is almost always negative and its absolute value is small compared to σ_{dw}. In other words, when the HT estimator is appropriate it is essentially using the variance of the constructed population based on its weights to get its estimate of variance.

The only difference between populations A and B is that a constant has been added to the y value of each unit. Now if the sample weights allow us to make a good guess for the population in the first case what goes wrong in the in the second case to cause the HT estimator to preform so poorly? To see the problem consider the following.

In the HT estimate the sum of the weights in the sample almost never equal N, the population size. Given a sample in population B the HT estimate is

$$\sum_{i=1}^{50} w_i y_i = \sum_{i=1}^{50} w_i y_i' + 400 \sum_{i=1}^{50} w_i$$

where y_i' denotes the unit's corresponding value in population A and y_i its value in population B. Note the second term in the above equation is adding additional variablity to the HT estimator. In population B calculations show that the term γ_{dw} in equation 9 is positive and can be quite large. It is accounting for the extra variablity in the HT estimator in population B which results from that fact that here $y_i \propto x_i + 400$ and not x_i.

We note that Zheng and Little (2003) argued that when estimating a finite population total and when using a probability-proportional to size sampling design that a penalized spline, nonparametric, model based estimator generally outperformed the Horvitz Thompson estimator. Zheng and Little (2005) developed methods to estimate the variance of their estimator. Some related work can be found in Zheng and Little (2004)

The WDP weights only use the constraint that simulated complete copies of the population should have the correct population mean for x. This is a more robust assumption than the one which underlies the HT estimator. But to be fair to the HT estimator it should be remembered (as was pointed out by a referee) that it was developed with the limited goal of obtaining linear unbiased estimators of the population total. Today however its simplicity no longer seems so important when more complicated and efficient estimators are much easier to compute. The superior performance of the stepwise Bayes method here suggests that if one believes that they have a set of weights for the sampled units which sums to the population size and which yields a good guess for the population, then they should use the variance of their good guess for the population to construct an estimate of the variance of their estimate of the population mean rather than equation 8. This is particularly true for large surveys containing several y characteristics of interest. It would be very surprising if all of them satisfied the assumptions necessary to make equation 8 a good estimate of variance of a sample mean. Analogous to the observation in Royall and Cumberland (1981) and Royall and Cumberland (1985) that good balanced samples (the sample mean is close to the population mean) can lead to improved performance one should base their inference on simulated complete copies of the population which incorporate the available prior information contained in the auxiliary variables.

7 Examples

We believe that standard design based theory over emphasizes the role that the selection probabilities should play in making inferences after the sample has been observed. In this section we consider examples that show how the WDP can make use of objective prior information after the sample has been selected.

7.1 A simulation study

To further understand how using the stepwise Bayes weights in the WDP can work we did a simulation study. We constructed a population with 2,000 units and a single auxiliary variable, x. This variable was a random sample from a gamma distribution with shape parameter 5 and scale parameter 1. The conditional distribution of y_i given x_i was normal with mean $100 + (x_i - 8)^2$ and standard deviation 20. The correlation for the resulting population was -0.38. We denote this population by quad. Clearly this is a toy example and the particular form of the relationship between x and y is not important to the WDP methods beyond the fact that x does contain some information about y. In what follows we will compare WDP estimators to two standard methods under four different sampling plans.

To construct the CPP we assumed that the x values for the population are known and we use them to construct three strata after the sample has been observed. These strata will not be constructed in the usual way. We did this to underplay the usual role of the design and to emphasize the robustness of our approach against the choice of design. We will have a sample size of $n = 60$ and we will construct three post-strata. Let $x_{[1]} < x_{[2]} < \cdots < x_{[60]}$ be the order statistic of the x values in the sample.

Let q_{20} and q_{40} be the population quantiles of $x_{[20]}$ and $x_{[40]}$ respectively. Then the CPP assumes that the total probability assigned to the units in the sample with the 20 smallest x values must be q_{20} and the total probability assigned to the next 20 smallest must be $q_{40} - q_{20}$. In other words we break the sample into three equal groups and use the information in the x values to get the appropriate population size of the corresponding strata. In addition the CPP assumes that the probabilities assigned to the sample must satisfy the population mean constraint for x.

The resulting WDP will be compared to two standard frequentist methods. The first is the post-stratified estimator which makes use of the same strata information as the CPP. The second is the usual regression estimator which assumes that the population mean of x is known. Although the regression estimator is not really appropriate for population quad it is included as a comparison. When computing 95% confidence intervals for the population total both frequentist methods will assume simple random sampling even when different sampling designs were used. We will denote these two estimators by STR and REG respectively.

The first sampling design was simple random sampling without replacement. For the second we generated a set of sampling weights by taking a random sample of 2,000 from a gamma distribution with shape parameter 5 and scale parameter 1. We then added 5 to each value to get the vector, v say. Note the values of v and y are completely independent. We then used approximate $pps(v)$ where at each step the probability that a unit is selected is proportional to its v value and depends only the unselected units remaining in the population. We call this the Random Weights design. For the third design we used approximated $pps(x)$. For the fourth we found the linear function, say l, which maps the range of y onto the the interval $[1, 2]$. We then used approximate $pps(l(y))$ as the sampling design. We call this the y Dependent design. In this design the selection probabilities depend weakly on the y values and units with large y values are more likely to be selected than those with small values of y. In particular the unit with the largest y value is twice as likely to be selected as the unit with the smallest y value. Clearly the Random Weights design and the y Dependent design are not standard designs and would never be used in practice. They were included to emphasize our belief that in many cases given a sample a good estimate does not depend on how the sample was selected.

For each design we took 500 samples of size 60 and computed the point estimate, its absolute error, the length of its interval estimate and whether or not it contained the true parameter value. The results are given in Table 2.

Remember that in this example the WDP is using information from both the post-stratification and knowing the population mean of x while STR just uses the first and REG just uses the second. Under SRS and the Random Weights design all four methods preform about the same. For the other two designs WDP does the best. Over all four designs its frequency of coverage is closest to the nominal level of 0.95. Using the constraint involving the population mean of x allows it to correct for some of the bias introduced by the sampling plans that STR cannot do. However this constraint can only do so much. If in the y dependent design the range of l was $[1, 4]$ then WDP's average absolute error is 4.5% better then that of STR and the frequency of coverage on the 0.95 nominal intervals were 0.86 and 0.80 respectively. There is just not enough information in x to correct for this much selection bias.

For each design we have included the average of the smallest and largest values of the parameter values defining the WDP which in this case must sum to 60. We see the range is largest for pps(x).

In the simulations we also used the WDP to construct 0.95 credible intervals for the population median of y. For the four designs its respective frequency of coverage was 0.956, 0.950, 0.952 and 0.930.

We did another simulation study where x was generated in the same way but now the conditional distribution of y_i given x_i was normal with $60 + x_i$ and standard deviation $2\sqrt{x_i}$. The correlation between x and y was 0.46. Under all four designs the performances of the point estimators were very similar. The WDP intervals tended to be a bit longer than the rest but over the four designs its average frequency of coverage for the population total was 0.949. Under the y Dependent design its frequency of coverage for the population total was 0.934 while for STR and REG the corresponding coverages were 0.896 and 0.886. Its average frequency of coverage for the population median of y was 0.942.

A frequentist could argue that this is an unfair example since the regression estimator does not make much sense for this population and of course they would be right. If for this problem you assumed a quadratic relationship between y and x and if you assumed that the first two population moments of x were known then the resulting regression estimator would out perform the WDP. In Lazar et al. (2008) there is such an example. Moreover, they show that including a constraint for the second moment of the CPP will hardly change the behavior of the resulting estimates. Hence, when there is good prior information about the model relating x and y this should be used in the analysis. When such prior information is not available we believe the WDP does have certain advantages even though it may not yield dramatic improvements over standard methods. It uses only objective prior information and makes no model assumptions about how the characteristics of interest and the auxiliary variables are related. It can correct for a slight dependency of the selection probabilities on the characteristic of interest. Although the sampling design plays no explicit role in its calculation, information which is often incorporated in the design can be reformulated as a constraint and be used when defining the CPP. Given a sample, inferences based on the WDP use many simulated complete copies of the population which on the average are consistent with the prior information. This makes makes it straightforward to estimate parameters other than a population mean or total.

7.2 Stratification and estimating the median

In many applications only a few observations, sometimes only two, are taken from each stratum. For such problems finding a good confidence interval when estimating the population median can be difficult. Next we will compare the standard method, see for example section 5.11 of Särndal et al. (1992), with the WDP. We will assume simple random sampling without replacement within strata.

For definiteness, assume we have L strata and stratum j contains N_j units. Let $N = \sum_{j=1}^{L} N_j$ be the total size of the population. Assume that two observations are taken from each stratum. Then the weight assigned to each sampled unit is one-half of the stratum size from which it was selected. The standard method uses these weights to find its confidence interval.

For this scenario the usual Polya posterior is applied within each stratum, independently across strata. Alternatively, this can be thought of as a CPP where the amount of probability assigned to the two sampled units in stratum j must sum to N_j/N. If $p_j = (p_{j,1}, p_{j,2})$ represents the probability assigned to the two sampled units from stratum j then under the CPP $E(p_j) = (N_j/(2N), N_j/(2N))$. Recalling the notation from section 5 we see that under the WDP the weight assigned to each of the two sampled units in stratum j is $(LN_j)/N$. Recall that simulating complete copies of the population using the WDP means that individual simulated copies will almost certainly not satisfy the constraints however the constraints will be satisfied when we average over all simulated copies. At first glance this might seem like a bad idea but we will see that when estimating the population median interval estimates based on the WDP behave better than the standard intervals which are too short. We shall see that the extra variability present in the WDP yields longer intervals with better frequentist properties.

The stratified populations we considered were constructed as follows. The strata sizes were a random sample from a Poisson distribution with parameter $\lambda = 100$.

The strata means were a random sample from a normal population with the mean $\mu = 150$ and with either a standard deviation of $\sigma = 10$ or $\sigma = 20$. The strata standard deviations were a random sample from a gamma distribution with scale parameter one and shape parameter $\gamma\sigma$ with either $\gamma = 0.10$ or $\gamma = 0.25$. We constructed two versions of each of the four types, one with 20 strata and the other with 40 strata. For each of the eight populations we took 500 samples where each sample consisted of two observations selected at random without replacement from each stratum. For each sample we compared the standard approach with estimates based on the WDP. The results can be found in Table 3. We only present the results for the 20 strata populations because the results for the 40 strata population are similar. Both methods are approximately unbiased and the point estimate based on the WDP seems to do just a bit better. But the confidence intervals produced by WDP are clearly superior. Even though in one case the WDP intervals are clearly too long its overall performance is much better than the standard intervals.

What causes the poor performance of the WDP intervals in the one case? Additional simulations indicate that when the strata means vary widely and the strata variances tend to be relatively small then the WDP intervals will tend to be too long. In our simulations the case with $\sigma = 20$ and $\gamma = 0.10$ leads to a population with such strata. When the sample size was increased to four units per stratum the difference between the two methods is not so dramatic but the story remains much the same. The standard intervals tend to be to short and under cover while the WDP intervals are longer and tend to over cover.

Clearly the choice of a good method for constructing a confidence interval depends not only on the size of the intervals it produces and but on the probability with which those intervals fail to include the true but unknown parameter value. Cohen and Strawderman (1973) and Meeden and Vardeman (1985), among others, have explored the question of admissibility for confidence intervals. Although the results given there are not directly applicable to our case the second paper shows that in some situations certain Bayes procedures can yield almost admissible procedures. These type of arguments along with the fact that the standard interval is way too short gives some circumstantial evidence, we believe, that the WDP intervals in this example are not outrageously too long. To sum up, we believe that in the important special case when the sample sizes are two and the strata are not dramatically different the WDP intervals seem to be a serious competitor for the standard intervals.

7.3 Integrated public use microdata series

The Minnesota Population Center (MPC) is an interdepartmental demography research group at the University of Minnesota. A major goal of the MPC is to create databases and statistical tools which can be utilized in the study of economic and social behavior. One database of interest is the Integrated Public Use Microdata Series (IPUMS), which is a consolidation of U.S. censuses and other national surveys from 1850-present (Ruggles et al., 2004). The word **microdata** is applied in this context because each row of an IPUMS dataset corresponds to one individual or one household; such low-level of detail may be contrasted with a typical Census Bureau publication or online summary table, in which a preset geographic specific tabulation (geography can be the entire country, states, counties, census tracts etc.) of the microdata is given to the data user.

One dataset which offers a rich array of numerical variables is the 2005 American Community Survey (ACS). This Census Bureau product is a large sample survey, and the Census Bureau does not know the true population means for the variables. To conduct simulations with the 2005 ACS, the sample played the role of the population. More specifically, the full population was assumed to be a set of 3,579 Minneapolis residents who are of working age (between 25 and 75), and who earn a yearly wage between \$20,000 and \$120,000. For our purposes the two variables of interest were:

- *inctot*. Total pre-tax income from 2004.
- *sei*. The Duncan Socioeconomic Index. Created in the 1950's, this is a numerical variable which attempts to rate the prestige associated with an individual's occupation. The range of this variable is $[1, 100]$.

For our simulations we set $y = \log(inctot)$ and $x = sei$. The correlation between y and x is 0.398 and we assume that the mean of x is known. For estimating the population mean of y we considered the estimator based on the WDP and the regression estimator. We used two different designs: simple random sampling and approximate pps(x). In each case we took 300 samples of size 30. The results are given in Table 4. We see that although the two methods are comparable the WDP clearly gives the better intervals.

8 Final remarks

The construction of weights in survey sampling is often more of an art than a science. This is one possible conclusion that can be drawn from the recent paper of Gelman (2007) and the accompanying discussion. He argues for a Bayesian approach to constructing weights using regression models which relate the characteristic of interest to auxiliary variables. Here we argued for a stepwise Bayes approach which will make use of the information present in the auxiliary variables without assuming a model relating the characteristic of interest to the auxiliary variables. The resulting weight for a unit in the sample can be given the usual interpretation as the number of units in the population which it represents.

A frequentist weight, say w_i, is the inverse of an inclusion probability, and this number represents the number of units in the population represented by a particular unit in the sample. So $w_i \geq 1$ for all i and $\sum_{i \in s} w_i \approx N$. In section 6 we saw that for the Horvitz-Thompson estimator the sum of the weights of the units usually fails to equal the population size which can result in a poor estimator except in very special circumstances. Another problem with frequentist weights is that they are often adjusted—after the sample is collected—to ensure that the frequentist estimates are in agreement with prior information about the population (Kostanich and Dippo, 2002). After making adjustments, the weights may be rescaled so that they sum to a population total. However, the adjusted frequentist weights no longer depend just on the sampling design and they no longer represent inverses of inclusion probabilities. The intuition behind frequentist weights is therefore somewhat confusing. Before adjustments, frequentist weights are functions of the design; but after adjustments, they are now functions of the design and other prior information, which may or may not be related to the design.

Bayesians think of estimation in survey sampling as a prediction problem. Their predictions are based on an assumed model which can lead to weights being assigned to the units in the sample. See for example the aforementioned Gelman (2007) and Little (2004). As noted by a number of authors (Pfeffermann, 1993) performing a weighted analysis for a model using inverses of the inclusion probabilities can protect the sampler from model misspecification. Moreover in certain situations the two approaches may lead to similar results.

Recently, Rao and Wu (2010) have developed methods which use a pseudo empirical likelihood approach and base their inferences on Dirichlet posterior distributions. The resulting procedures, although formally somewhat similar to some discussed here, use prior information in a different way. For them much of the prior information must be filtered through the design while we believe that prior information which is often included in the design can be used directly to generate good posteriors. For better or worse we are closer to the classical Bayesian scenario where the posterior distribution does not depend on the sampling design.

Here we have focused on using the CPP to generated a set weights based on the sample and prior information and then making our inferences using the WDP based on these weights. Strief (2007) considered examples where the weights generated by the CPP were instead used in the appropriated frequentist formulas to get an estimate of variance and noted that their performance was similar to standard methods. Alternately one could imagine basing their inferences on the WDP but using frequentist weights, say generated by calibration methods (Särndal and Lundström, 2005), instead. Although this deserves further study it is our expectation that such approaches should lead to inferential procedures with good frequentist properties.

In the design based approach consistency is an important property for an estimator to possess. For an important special case when the design is SRS the CPP estimators are consistent. This is demonstrated in Geyer and Meeden (2011).

Just as the CPP does, the WDP also has a stepwise Bayes justification. (For more details see Strief (2007).) The weights used in the WDP have a consistent formulation and interpretation. They are always a posterior expectation and always sum to the population size. They represent the average number of times that each unit in the sample appears in a simulated, completed copy of the population under the CPP. This average is with respect to the uniform distribution over all possible copies of the population which just contain the units in the sample and which satisfy the given constraints. These weights depend only on the same kinds of objective prior information about the population which are often used to define and adjust frequentist weights. This allows them to incorporate prior information without explictly specifying a prior distribution.

In most cases the weight assigned to a unit in the sample will depend on the other units in the sample. We have argued that after the sample has been selected one should argue conditionally. That is, given the sample the weights should depend on all the available prior information about the population but not on how it was selected. (We are assuming that the person selecting the sample and the analyst are one in the same.) Any procedure constructed in this manner should preform well for a variety of sampling designs. For any procedure, be it either frequentist, Bayesian or stepwise Bayes this is the litmus test: it should be evaluated by how it behaves under repeated sampling from the design of interest.

To implement the methods discussed here one first needs to use the CPP to computed the weights for the observed sample. Then one needs to use the weights in the WDP to simulate complete copies of the population. The first step is the more difficult although the software package *polyapost* makes it relatively straightforward for anyone familiar with R. Once the weights are known it is easy to simulate from the WDP in many computer packages. This makes our approach more practical for survey datasets (like IPUMS) which are presented with the weights attached and are used by multiple researchers. A more serious limitation is that we have only considered simple single stage sampling designs. More work needs to be done to extend these methods to more complicated multi-stage designs. If the underlying constraints are selected wisely the resulting procedures can have good frequentist properties for a variety of sampling designs. These stepwise Bayes weights can be thought as our best guess for the unknown population given the sampled units and our prior information.

References

Binder, D. (1982). Non-parametric Bayesian models for samples from a finite population. *Journal of the Royal Statistical Society, Series B*, 44:388–393.

Booth, J. G., Bulter, R. W., and Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89:1282–1289.

Cohen, A. and Strawderman, W. (1973). Admissible confidence interval and point estimation for translation of scale parameters. *Annals of Statistics*, 1:545–550.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22:153–188.

Geyer, C. and Meeden, G. (2011). Asymptotics for constrained dirichlet distributions. submitted.

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

Gross, S. (1980). Median estimation in survey sampling. In *Proceedings of the section of section on survey research methods*, pages 181–184. American Statistical Association.

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55:159–167.

Hu, M., Zhang, F., Cohen, M., and Salvucci, S. (1997). On the performance of replication-based variance estimation methods with small number of psus. In *Proceedings of the Survey Research Methods Section*. American Statistical Association.

Kostanich, D. L. and Dippo, C. S. (2002). Design and methodology: 63rv. Technical report, The U.S. Census Bureau and The Department of Labor Statistics.

Lazar, R., Meeden, G., and Nelson, D. (2008). A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Survey Methodology*, 34:51–64.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite poplation sampling. *Journal of the American Statistical Association*, 99:546–556.

Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16:1684–1695.

Meeden, G. and Vardeman, S. (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association*, 80:465–471.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61:317–337.

Rao, J. N. K. and Wu, C. (2010). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, 72:533–544.

Royall, R. and Cumberland, W. (1981). An empircal study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 71:657–664.

Royall, R. and Cumberland, W. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80:355–359.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9:130–134.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P. K., King, M., and Ronnander, C. (2004). Integrated public use microdata series: Version 3.0 [machine-readable database]. University of Minnesota.

Särndal, C.-E. and Lundström (2005). *Estimation in surveys with nonresponse.* Wiley, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling.* Springer, New York.

Strief, J. (2007). *Bayesian sampling weights: toward a practical implementation of the Polya posterior*. PhD thesis, University of Minnesota.

Team, R. D. C., editor (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, www.R-project.org.

Zheng, H. and Little, R. (2003). Penalized spline model-based estimation of finite population total from from probability-proportional-to-size samples. *Journal of Official Statistics*, 19:99–117.

Zheng, H. and Little, R. (2004). Penalized spline nonparametric mixed-model for inference for finite population means from two-staged samples. *Survey Methodology*, 30:209–218.

Zheng, H. and Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21:1–20.

Table 1: Results for populations A and B based on 500 samples of size 50. The NHT estimator is the HT estimator renormalized so that the weights sum to the population size, $N = 2,000$. The nominal coverage for each method is 0.95

Population	Method	Ave. abs err	Ave. len	Freq of coverage
A	HT	4,628	21,898	0.940
B	HT	8,965	43,914	0.960
A & B	WDP	4,706	24,381	0.960
A	NHT	5051	21,897	0.896
B	NHT	5051	43,919	0.998

Table 2: Simulation results for population quad discussed in section 7.1 for 500 random samples of size 60 for four different sampling plans. The true population total was 227,923.0. The nominal coverage for each method is 0.95

Method	Ave. value	Ave. err	Ave. len	Freq of coverage
		SRS		
STR	227,856.1	4,165.0	21,332.1	0.950
REG	227,602.1	4,302.7	21,300.3	0.944
WDP	227,546.9	4,190.6	23,029.7	0.958
Ave. min and max of WDP parameters were 0.658 and 1.580.				
		Random Weights		
STR	227,976.5	4,371.2	21,254.1	0.938
REG	227,715.5	4,462.2	21,305.9	0.934
WDP	227,721.2	4,420.6	22,901.4	0.950
Ave. min and max of WDP parameters were 0.651 and 1.583.				
		pps(x)		
STR	225,295.8	5,228.9	23,008.4	0.916
REG	224,207.2	5,611.2	21,780.3	0.878
WDP	227,471.1	4,919.2	22,706.6	0.936
Ave. min and max of WDP parameters were 0.374 and 3.024.				
		y Dependent		
STR	231,590.0	5,229.0	21,170.8	0.892
REG	231,424.4	5,143.4	21,127.9	0.902
WDP	231,139.1	4,967.6	22,867.0	0.938
Ave. min and max of WDP parameters were 0.660 and 1.643.				

Table 3: Simulation results from 500 stratified random samples of size two within each strata from populations with 20 strata. The nominal coverage for each method is 0.95

Method	Ave. value	Ave. err	Ave. len	Freq of coverage
	$\sigma = 10$ and $\gamma = 0.10$			
Stand	148.40	2.37	8.30	0.808
WDD	148.39	2.22	12.20	0.95
	$\sigma = 10$ and $\gamma = 0.25$			
Stand	144.28	5.70	20.59	0.834
WDD	144.18	5.41	28.38	0.950
	$\sigma = 20$ and $\gamma = 0.10$			
Stand	152.75	3.02	10.52	0.828
WDD	152.61	2.78	22.88	0.996
	$\sigma = 20$ and $\gamma = 0.25$			
Stand	155.94	6.72	23.17	0.826
WDD	155.89	6.35	34.96	0.962

Table 4: Simulation results from 300 random samples of size 30 from the IPUMS population. The nominal coverage for each method is 0.95

Design	Method	Ave.	Ave.	Freq of
		err	len/2	coverage
SRS	Reg	0.052	0.128	0.943
	WDP	0.052	0.138	0.947
PPS(x)	Reg	0.062	0.132	0.897
	WDP	0.066	0.133	0.937