Ordered Designs and Bayesian Inference in Survey Sampling

Glen Meeden*
School of Statistics
University of Minnesota
Minneapolis, MN 55455
glen@stat.umn.edu

Siamak Noorbaloochi†
Center for Chronic Disease Outcomes Research
Minneapolis VA Medical Center
Minneapolis, MN 55417
and Department of Medicine
University of Minnesota
Siamak.Noorbaloochi@med.va.gov

March 10

*Research supported in part by NSF Grant DMS 0406169
†Research supported in part by VAHSR&D Grant IIR 07-229

Ordered Designs and Bayesian Inference in Survey Sampling

Abstract

Many sampling designs, such as simple random sampling without replacement, can in principle be extended in a natural way so that the units continue to be selected until the population is exhausted. These designs impose a random order on the population. Here we show how such ordered designs can be used to define prior distributions over the population. For such priors the Bayesian analysis uses information that in standard frequentist methods is incorporated in the sampling design. The resulting methods will often have good frequentist properties.

Key Words and phrases: finite population sampling, sampling designs Polya posterior and noninformative Bayesian.

AMS 1991 Subject Classification: Primary 62D05; Secondary 62C10

1 Introduction

Godambe (1955) noted at the beginning of this influential paper that many sampling designs can be thought of as being defined conditionally on the order that the units appear in the sample. He then suggested that from a theoretical perspective it is convenient to ignore this fact and just consider the unordered sample (where the order is ignored). (For samples of size n there are $n!$ possible ordered samples corresponding to each unordered sample.) Murthy (1957) showed that for any estimator which depends on the ordered samples there exists an estimator which only uses the unordered samples which has the same expectation but smaller variance except when the two estimators are the same. This application of the Rao-Blackwell theorem was also discussed by Pathak (1961). For these reasons most sampling theory has concentrated on unordered designs. Raj (1956) is one example where the order in which the sample was drawn was considered.

The design probabilities play a fundamental role in standard frequentist theory. Since units in the sample are assumed to be observed without error the only randomness in the model comes from them and it is upon these probabilities that the frequentist properties of estimators are based.

In the Bayesian approach to statistical inference the posterior distribution summarizes the information about a parameter. This distribution depends on a probability model and a prior distribution and is conditional on the observed data. In finite population sampling the unknown parameter is just the entire population and a prior distribution must be specified over all possible values of the units in the population. Given a sample the posterior is just the conditional distribution of the unobserved units given the values of the observed units computed under the prior distribution for the population. This posterior does not depend on the sampling design used to select the sample. The Bayesian approach to finite population sampling was elegantly described in the writings of D. Basu. For further discussion see his collection of essays in Ghosh (1988). A problem with the Bayesian approach is that it can be difficult to find prior distributions which make use of available prior information about the population.

Once a sampling design has been selected, but before the sample is implemented, the actual units that will appear in the sample are unknown. This suggests, from a Bayesian perspective, that the design could be considered as part of the prior distribution. Here we will consider prior distributions which are defined in two steps. First, we randomly assign an order to the

units in the population. Then conditionally on a given order we define a distribution for the possible values of the units. This procedure allows our prior to capture some of the information that is present in designs that select units sequentially. We will call such designs ordered designs. Our main goal is to study Bayesian models where an ordered design is used in the first step of the process of defining a prior.

In section 2 we introduce a new way to define prior distributions in finite population sampling. In section 3 we use a stepwise Bayes argument to prove an admissibility result for this new method of defining prior distributions. In section 4 we consider two examples. In the first we study smooth populations where units with labels close together tend to be more similar than units whose labels are far apart. In the second we consider a situation where the sampling design depends on the values of the characteristic of interest. In section 5 we conclude with a brief discussion.

2 A family of priors

Before discussing our method for defining prior distributions we need to introduce some notation. We need to warn the reader that our notation is not standard.

2.1 Notation

The population size is N and $\Lambda = \{\alpha, \beta, \ldots, \tau\}$ is a set of N labels which identify the units in the population. We let λ denoting a typical label. Let $y = \{y_\lambda : \lambda \in \Lambda\}$ denote a typical unknown set of population values. Let $b = (b_1, b_2, \ldots, b_k)$ be a vector of known positive real numbers. We assume for technical convenience that the only possible values for each y_λ is some b_j. Hence our parameter space will be

$$\begin{aligned}\mathcal{Y}(b) =& \{y : \text{such that for each } \lambda \in \Lambda \text{ there exists} \\ & \text{some } j = 1, 2, \ldots, k \text{ for which } y_\lambda = b_j\} \end{aligned} \tag{1}$$

Λ is an unordered set since the population labels have no order. However order will be important for us. This is achieved by letting $u = (\alpha, \beta, \ldots, \tau)$ denote the labels in some fixed order. Then y_u denotes y arranged in this standard default order. So when we write y order does not matter while it

does matter in y_u. More generally $\{\}$ applied to an ordered set means that order no longer is taken into account. So we can write $y = \{y_u\}$.

If π is a permutation of $1, 2, \ldots, N$ we let $\pi(u)$ be the permutation π applied to u to give a new order for the labels. $y_{\pi(u)}$ denotes the values of y arranged in the order determined by $\pi(u)$. For simplicity we write y_π instead of $y_{\pi(u)}$. Let Π be the set of all possible permutations of $1, 2, \ldots, N$. Since order will matter for us another space of interest is

$$\mathcal{Y}(b, \Pi) = \{(\pi, y_\pi) : \text{ where } y_u \in \mathcal{Y}(b) \text{ and } \pi \in \Pi\} \tag{2}$$

For each fixed y_u this set will contain $N!$ points, one for each possible permutation π. For each π the point (π, y_π) consists of the permutation along with the order of y_u under this permutation.

We let π_n denote the first n values of π and π_{n+} its remaining $N - n$ values. Similarly we let y_{π_n} be the first n values (in order) of y_π and $y_{\pi_{n+}}$ be the remaining $N - n$ values (in order).

Suppose for a moment that the sampling design is simple random sampling without replacement (srs) with a sample size of n. In this case the labels of the units to be observed will be selected one at a time until a sample of size n is drawn and their order of selection will be noted. At this point, before the the y values of the units selected are observed, we can imagine continuing the sampling procedure until all the units from the population have been selected and given an order. This is just a thought experiment and is not something that would be implemented. However we see that the srs design can be extended in a natural way to define a probability distribution on Π. When the design is srs the resulting distribution is just the uniform distribution on Π. Before the labels are selected and the characteristic of interest observed we can think of both π and y_π as unknown. Observing the data results in partial information about both of them. From the Bayesian perspective this means we could define a joint prior distribution over the pair on the space $\mathcal{Y}(b, \Pi)$. In the next section we will see how this can be done.

Other ordered designs can be given the same treatment. For example if there is a real valued auxiliary variable and at each stage a unit is selected at random with probability proportional to the size of its auxiliary variable (pps) then this also defines a probability distribution on Π.

2.2 Defining a prior

If we take $\mathcal{Y}(b, \Pi)$ as the space of possible values for the unknowns then our prior distribution can be defined on this space. The joint distribution for (π, y_π) can be written as a marginal for π and a conditional for y_π given π. Our key idea is that we can use the probability distribution on Π, coming from the ordered design, as the marginal for π. Then given an order, π, it remains to define the conditional distribution of y_π given π. Under this setup our prior distribution will be of the form

$$p(\pi, y_\pi) = p(\pi)\, p(y_\pi \mid \pi) \tag{3}$$

Then the prior distribution for y is just the marginal distribution for y under the above model. As far as we know such priors have not been considered in the literature. We will see that this has some interesting consequences.

In the following sometimes order will matter as in the previous equations and at other times it will not. Our notation will try to make this clear.

Let

$$s = (\lambda_1, \lambda_2, \ldots, \lambda_n)$$

denote the labels of a sample of size n in order that they were selected. Let

$$y_s = (y_{\lambda_1}, y_{\lambda_2}, \ldots, y_{\lambda_n})$$

denote the corresponding y values of the variable of interest. Note order matters here. As noted earlier $\{s\}$ will denote the unordered version of s and $\{y_s\}$ its unordered set of y values.

Even though we are using an ordered design to define a prior distribution we will see that the resulting estimators will depend on the observed data only through $(\{s\}, \{y_s\})$. This is because any admissible estimator must be a function of the sufficient statistic.

Let (π, y_π) be consistent with the unordered values of the observed sample (s, y_s). This means that $y = y_\pi$ for some π where

$$\{\pi_n\} = \{s\} \quad \text{and} \quad \{y_{\pi_n}\} = \{y_s\}$$

Note for srs and pps it is true that

$$p(\pi) = p(\pi_n)\, p(\pi_{n+} \mid \{\pi_n\}) \tag{4}$$

So we will assume that our marginal prior distribution on Π satisfies this equation for our fixed sample size n and each choice of π.

For each π and the fixed sample size n we will assume that

$$p(y_\pi \mid \pi) = p(y_{\pi_n} \mid \pi_n)\, p(y_{\pi_{n+}} \mid \{y_{\pi_n}\}, \pi_{n+}) \tag{5}$$

This assumption breaks $p(y_\pi \mid \pi)$ into two factors. The first factor states that given π the distribution of the first n units depends only on π_n and not on anything that comes later. The second factor states that the distribution of the last $N-n$ units can depend on the first n units only through their unordered values. This may seem like a strong assumption but we will see that it is satisfied for some models which are useful in survey sampling.

With these two assumptions we can write the joint distribution of (π, y_π) as

$$p(\pi, y_\pi) = p(\pi_n)\, p(y_{\pi_n} \mid \pi_n)\, p(\pi_{n+} \mid \{\pi_n\})\, p(y_{\pi_{n+}} \mid \{y_{\pi_n}\}, \pi_{n+}) \tag{6}$$

We now wish to find

$$p(\pi_{n+}, y_{\pi_{n+}} \mid \{s\}, \{y_s\}) = \frac{p(\{s\}, \{y_{\pi_n}\}, \pi_{n+}, y_{\pi_{n+}})}{p(\{s\}, \{y_s\})} \tag{7}$$

Let $\{s\}$and $\{y_s\}$be fixed and let B be the set of (π, y_π) consistent with them. That is

$$B = \{(\pi, y_\pi) : \text{where } \{\pi_n\} = \{s\} \text{ and } \{y_{\pi_n}\} = \{y_s\}\}$$

Then

$$\begin{aligned}
p(\{s\}, \{y_s\}) &= \sum_{(\pi, y_\pi) \in B} p(\pi_n)\, p(y_{\pi_n} \mid \pi_n)\, p(\pi_{n+} \mid \{\pi_n\})\, p(y_{\pi_{n+}} \mid \{y_{\pi_n}\}, \pi_{n+}) \\
&= \sum_{(\pi, y_\pi) \in B} p(\pi_n)\, p(y_{\pi_n} \mid \pi_n) \Bigg(\sum_{\pi_{n+}:\{\pi_n\}=\{s\}} p(\pi_{n+}) \Bigg[\\
&\qquad \sum_{y_{n+}:\{y_n\}=\{y_s\}} p(y_{\pi_{n+}} \mid \{y_{\pi_n}\}, \pi_{n+}) \Bigg] \Bigg) \\
&= \sum_{(\pi, y_\pi) \in B} p(\pi_n)\, p(y_{\pi_n} \mid \pi_n)
\end{aligned}$$

since the sum within the brackets is 1 which then yields that sum within the parentheses is 1. From this it follows easily that

$$p(\pi_{n+}, y_{\pi_{n+}} \mid \{s\}, \{y_s\}) = p(\pi_{n+} \mid \{\pi_n\}) p(y_{\pi_{n+}} \mid \{y_{\pi_n}\}, \pi_{n+}) \tag{8}$$

This has the same form as the prior in equation 3 with each factor updated in a sensible manner. Note that of the two factors in equation 5 defining $p(y_\pi|\pi)$ the first, $p(y_{\pi_n} \mid \pi_n)$ plays no role in the posterior. Hence there is no need to specify it. We just assume that it has been done in some consistent manner. Although we cannot find explicit expressions for the corresponding point estimate of the population total along with its posterior variance they often can be found easily through simulation. Generating simulated copies of $y_{\pi_{n+}}$ from this distribution can be done in two steps. First we simulate a set of values for π_{n+} from the first factor. Then we plug these values into the second factor and use it to get a set of simulated values for the unseen members of the population.

2.3 An example

Consider a population with y, the characteristic of interest, and x, an auxiliary variable. We assume that x and y are positively correlated and that units with x values close together tend to be more similar than units whose x values are far apart. We will now see how the priors discussed here can capture that information without making any model assumptions about how x and y are related. In addition we assume that the x values are known for all the units in the population.

We assume the design is srs so $p(\pi)$ is the uniform distribution over the $N!$ possible permutations. Hence $p(\pi_{n+} \mid \{s\})$ is just the uniform distribution over the $(N-n)!$ units in $\{s'\}$. It remains to define the second term of equation 8, $p(y_{\pi_{n+}} \mid \{y_{\pi_n}\}, \pi_{n+})$. To this end, if α and β are two labels belonging to Λ, we assume that x_α and x_β are distinct and that there is a distance function, $d(x_\alpha, x_\beta) > 0$, which measures the distance between them.

Suppose now we have observed our sample $(\{s\}, \{y_s\}, \{x_s\})$ where $\{x_s\}$ are the values of the auxiliary variable for the units in the sample. Given the sample these values can be assumed to be fixed. To describe our posterior in equation 8 for this special case we imagine two urns.

Into the first urn we place the n sampled units. Attached to each unit is their known x value and their observed y value. Into the second urn we place the $N-n$ unsampled units. Attached to each of them is their known x value. To get a simulated copy of the entire population we proceed as follows. We randomly select a unit from the second urn using srs. Denote this unit by x_α. The next step is to select a unit from the sampled units in the first urn and assign its y value to the unit α selected from the second urn. To select

the unit from the first urn we use pps sampling where the weight assigned to unit λ_i is $d(x_{\lambda_i}, x_\alpha)$ for $i = 1, 2, \ldots, n$. Then the unit α, with its simulated y value and known x value, is placed in the first urn and the unit selected from the first urn which was used in constructing this simulated value is also returned to the first urn. For the rest of the steps this unsampled unit with its simulated y value is treated exactly the same as the n original sampled units. This process is repeated with another unit being select from each urn and both placed in the first urn with the unsampled unit been given a simulated y value. This process will be continued until each unobserved unit has been assigned a y value. Thus we have constructed one simulated copy of the entire population given the sample using this predictive distribution for the unseen given the seen.

3 Admissibility

Consider the problem of estimating the population total, $t(y)$, under our scenario with the parameter space given in equation 1. Let δ denote an estimator of the population total. For a fixed parameter point y its expected loss is given by

$$E_y\big(\delta(\{s\}, \{y_s\}) - t(y)\big)^2 = \sum_{\pi \in \Pi} \big(\delta(\{s\}, \{y_s\}) - t(y)\big)^2 p(\pi_n)$$

which agrees with the standard formulation. Moreover, to prove admissibility for an estimator δ, it will be enough show that $p(y_\pi \mid \pi)$ is chosen properly.

Actually the posterior we described in section 2.3 is not a true posterior since we never specified $p(y_{\pi_n} \mid \pi_n)$. However it does have a stepwise Bayes justification which implies that it will yield admissible estimators. The rest of this section is a justification of the previous sentence.

In stepwise Bayes arguments a finite sequence of disjoint subsets of the parameter space is selected, where the order of the subsets is important. A different prior distribution is defined on each of the subsets. First, the Bayes procedure is found for each sample point that receives positive probability under the prior defined on the first subset. Next the Bayes procedure is found for each sample point which receives positive probability under the prior defined on the second subset and which was not considered in the first step. Then for the prior defined on the third subset the Bayes estimate is found for all sample points which are assigned positive probability and which

were not considered in the first two steps. This process is continued until all possible sample points have been considered. For a particular sample point the value of its stepwise Bayes estimate comes from the prior associated with the step where it was first assigned positive probability. It is the stepwise Bayes nature of the posteriors in equation 8 that explains their somewhat paradoxical nature. Given a sample each behaves just like a proper Bayesian posterior but the collection of possible posteriors that arise from all possible samples comes from a family of priors and not a single prior.

Theorem 1. *Consider the problem of estimating the population total with squared error as the loss function. Then the estimator arising from the posteriors described in section 2.3 is stepwise Bayes and hence is admissible when the sampling design is simple random sampling without replacement and the parameter space is* $\mathcal{Y}(b)$.

Proof. There are many stages in the proof and except for the first stage each have many similar steps. For each step within each stage we need to select a subset of $\mathcal{Y}(b)$ and define a prior distribution over the subset of the form given in equation 3.

At every step in each stage the distribution $p(\pi)$ is just the uniform distribution over the set of all possible permutations on a set of N objects.

In the first stage for each π $p(y_\pi \mid \pi)$ puts mass $1/k$ on each of the k choices of $y_\pi \in \mathcal{Y}(b)$ where every value of y_π is identically b_j for some choice of b_j. It is easy to see in this case that the only possible samples are those where all the observed y values are b_j for some j and the resulting estimator is Nb_j,

In the next stage we take care of all samples where the observed y values in the sample take on just two values. This is done in $\binom{k}{2}$ steps by considering samples where just (b_1, b_2) appear, then samples where just (b_1, b_3) appear and so on until the final step which considers samples where just (b_{k-1}, b_k) appear. In the next stage we take care of all samples where the observed y values take on just three values. We continue to the last stage where the samples take on $\min\{k, n\}$ different values.

We now consider some specific step in some specific stage. Let b^* be a subset of b of size $1 < m \leq \min\{k, n\}$. Without loss of generality we can assume that it is the first m elements of b. Let $\mathcal{Y}^*(b^*)$ be the subset of $\mathcal{Y}(b)$ where each of the values of b^* appear at least once. At this stage we only consider samples where the observed values in the sample consists of all the values in b^*. Samples where only some of these values appear have been

considered in an earlier stage. Since we have already defined $p(\pi)$ it only remains to define $p(y_\pi \mid \pi)$ for $y \in \mathcal{Y}(b^*)$.

We see from equation 8 that

$$p(y_{n+} \mid \{y_{\pi_n}\}, \pi_{n+}) \tag{9}$$

is what remains to be specified since $p(y_{\pi_n} \mid \pi_n)$ plays no role in the calculation of our estimate once $(\{s\}, \{y_s\})$ has been observed. Any choice of the above posterior is fine. In particular we can take the distribution described in section 2.3. Although there is no explicit expression for the resulting estimator it is stepwise Bayes and hence admissible. □

The proof of this theorem is quite similar to the stepwise Bayes proof of the admissibility of the usual estimator $(N/n)\sum_{i=1}^{n} y_{\lambda_i}$ of the population total. Using this theorem it is straight forward to show that our estimator given in section 2.3 is admissible when $\mathcal{Y}(b)$ is replaced by $(0, \infty)^N$. Details can be found in Ghosh and Meeden (1997).

It follows from a careful reading of the proof that any specification of

$$p(\pi_{n+} \mid \{s\}) \quad \text{and} \quad p(y_{y_{n+}} \mid \{y_s\}, \pi_{n+})$$

for every $(\{s\}, \{y_s\})$ will yield an admissible estimator. This is a reflection of the fact that admissibility is a rather weak property. However this flexibility should allow for the easy incorporation of various kinds of information into a Bayesian analysis which is not possible under a more standard Bayesian approach. Even if one cannot find explicit forms for the resulting estimator and its variance it will often be easy to simulate completed copies of the population using equation 8. In such cases one can find approximately the resulting estimator of many population quantities of interest. One can also find approximately the corresponding 0.95 Bayesian credible interval whose frequentist coverage will often be approximately 95%.

4 Two examples

4.1 Smooth populations

In section 2.3 we introduced a stepwise Bayes model for populations where units with labels that are close together tend to be more alike than units

whose labels are far apart. One situation where this can happen is when x and y are highly correlated.

Given the sample we select an unsampled unit from the second urn using srs. Suppose this unit has label α. We then select a unit at random from the first urn containing the sampled units using pps with weights proportional to

$$d(x_{\lambda_i}, x_\alpha) = exp - a|x_{\lambda_i} - x_\lambda| \qquad \text{for } i = 1, 2, \ldots, n \tag{10}$$

where $a > 0$ is specified by the statistician. The y value of this unit is assigned to the unit with label λ. This unit along with its assigned y value and known x is place in the first urn. The unit selected at random from the first urn is also returned to the first urn. This process is continued until all the unobserved units have been move into the first urn and have been assigned a y value. Clearly this defines a probability distribution for the "unseen" given the "seen" which makes use of the idea that we expect units whose x values are close tend to have similar y values.

To see how the resulting estimator could work in practice we constructed a population. We began by taking a random sample of size 200 from the uniform(10,14) distribution. Next we subtracted 12 from each of them and cubed the differences. Finally we added the same constant to each so that the minimum value of the set was 5. The resulting set was our population of x values. To get the y values we specified that for each λ the distribution of y_λ given x_λ was normal with mean x_λ and standard deviation 0.8. The correlation between x and y for the resulting population was 0.88 and the true population total of y was 2635.0.

For this population a frequentist could either stratify the population or use the regression estimator or perhaps do both. When using the regression estimator with stratification it is always a question of whether you should use one regression estimator for the entire population or use different regression estimators within the strata. Here we did both and denoted the two estimators by reg and regstr. Another frequentist estimator we considered was the method of collapsed strata. One might think that this could be a competitor to our procedure because in some sense both use more local information than the other estimators. This estimator was denoted by colspstrat while strat denoted the usual estimator based on the stratification. Finally, stB denotes our stepwise Bayes estimator with $a = 12$. We constructed 1000 simulated copies of the population when computing stB.

We considered two different sampling designs each using the ordered x

values. For the first design we constructed 4 strata of size 50 and selected 10 units from each strata using srs. For the second design we constructed 10 strata of size 20 and selected 4 units from each strata using srs. In each case we took 500 random samples. The results are given in Table 1.

Our stB procedure is the clear winner. As a point estimator it is beaten in only one case and then by just a bit and its interval estimates are superior across the board. Our stB model is using more information than the regression estimator since it assumes that the values of x are known for the population rather than just the mean. Also note that its posterior variance does not depend on any model assumptions about how the y and x values are related. Instead this comes from our stB posterior which just uses the assumption that units with x values which are close should tend to have similar y values.

We still need to discuss how the choice of $a = 12$ was determined. We see from equation 10 that as the value of a increases in the stB model the resulting simulated populations will have less variation. In particular we see that for a close to zero in the stb Model the resulting estimator should behave much like the sample mean. To see how sensitive the stB model is to the choice of a we generated another population using the same procedure that generated the one used in table 1. In this case the correlation between x and y was 0.89 and the population total was 2573.0. We then took 500 srs of size 50 and computed the usual estimator and the regression estimator. We also compute the stB Model estimator for five choices of a. The results are given in Table 2

We see for a large range of values for a the behavior of our stB point estimators do not vary much. For a smaller range of a their 0.95 credible sets will yield approximate 95% confidence sets for the population total. Choosing a either too small or too large results in simulated populations with too much variability or too little variability.

Clearly a good choice of a in equation 10 depends on the distance function. When the x_i's are real numbers and d is just ordinary distance we have found that the following rule of thumb seems to work reasonably well. We take $a \doteq N/(4R)$ where N is the population size and $R = \max_i x_i - \min_i x_i$. Note in our example $N = 200$ and $R = 4$ and our rule of thumb suggest that a choice of a close to 12.5 should work well. Next we will consider two more examples to see how our rule of thumb works.

In the first the mean of y is a cubic function of x. The values of x were a random sample of size 200 from the uniform distribution on the interval

[1, 27]. To get the y values we let

$$y_i = 4000 + 3x_i^3 - 117x_i^2 + 1140x_i + e_i$$

where the e_i's where independent normal random variables with mean 0 and standard deviation 300. $R = 26$ and $N = 200$ suggest that $a = 1.92$ should be a good choice. Based on the order of the x_i's we created 4 strata of size 50 each and then took 500 random samples where for each sample we selected 10 unites at random from each stratum. The results are given in Table 3. Again the stB procedure is the clear winner and our choice of $a = 0.25$ yields approximate 95% confidence intervals.

For the second example we consider a population consisting of the monthly means of daily relative sunspot numbers. The data is found in Andrews and Herzberg (1985). We consider the first 384 months where the x variable is just the labels $1, 2, \ldots, 384$. These numbers are highly cyclic in nature. We construct 6 ordered strata based on the labels. We then took 400 random samples where for each sample we selected 8 units at random from each stratum. Since in this case $R = N - 1 = 383$ our rule of thumb suggests that $a = 0.25$ should be a good choice. The results are given in Table 4. Again the stB procedure is the clear winner and our choice of $a = 0.25$ yields approximate 95% confidence intervals Note that we are not trying to model how neighboring units are related. We just need to make sure that the sample size is large enough to capture the structure of the population.

4.2 Designs depending on y

In certain situations, like observational studies, the design generating the units in the sample may not be under the control of the statistician. In some such cases it might be reasonable to assume that the probability a unit appears in the sample depends on its y value. Andreatta and Kaufman (1986) considered a problem in geology where the probability of discovering an oil field is proportional to the size of the field. In what follows we assume that this is the case, that is, the probability that a unit appears in the sample is roughly proportional to the size of its y value.

For these situations we can write the joint distribution for (π, y) as

$$\begin{aligned} p(\pi, y) &= p(y)\, p(\pi \mid y) \\ &= p(\pi_n \mid y)\, p(\pi_{n+} \mid \{s\}, \{y_s\})\, p(\{y_s\})\, p(y_{\pi_{n+}} \mid \{y_s\}, \pi_{n+}) \end{aligned} \tag{11}$$

Arguing as we did to get equation 8 we see that

$$p(\pi_{n+}, y_{\pi_{n+}} \mid \{s\}, \{y_s\}) = p(\pi_{n+} \mid \{s\}, \{y_s\})\, p(y_{\pi_{n+}} \mid \{y_{\pi_n}\}, \pi_{n+}) \qquad (12)$$

Hence to get admissible estimators we only need to specify the two posterior distributions in equation 12. They need to reflect the fact that we expect the unobserved units to tend to be smaller than the observed units in the sample.

It is not clear to us how to use this information when defining the first posterior in equation 12 so we will take it to be the uniform distribution over π_{n+}. However our assumption will come into play when we define the second posterior.

If y_λ appears in the sample then we are assuming that its selection probability was roughly proportional to y_λ. Standard design theory often recommends assigning a weight to a selected unit which is proportional to the reciprocal of its selection probability, $1/y_\lambda$. With this idea in mind for each unit y_α in the sample we define its weight to be

$$w_\alpha = n \frac{1/y_\alpha}{\sum_{\lambda \in s} 1/y_\lambda} \qquad (13)$$

We will now use these weights to define the second posterior in equation 12. We imagine two urns. The first contains the n units in the sample $s = (\lambda_1, \ldots, \lambda_n)$. The second urn contains the $N - n$ unsampled units. Each unit in the first urn has its weight, w_{λ_i}, attached. Note the total sum of the weights is n, the sample size. A unit, say λ_i is selected at random from the first urn with probability proportional to its weight. A unit is selected at random from the second urn and assigned the value y_{λ_i}. Both units are placed in the first urn and the selected unobserved unit with its imputed value is given a weight of 1. This process is repeated until all the unobserved units have been assigned a value. Thus, a simulated copy of the entire population is the result of Polya sampling from the first urn.

Let $p = (p_{\lambda_1}, \ldots, p_{\lambda_n})$ where p_{λ_i} is the proportion of units in a simulated copy of the population which take on the value y_{λ_i}. It is well known that if the population size is large compared to the sample then the distribution of p is approximately Dirichlet with parameter $w = (w_1, \ldots, w_n)$. So given a

sample the posterior mean of the population is approximately

$$\begin{aligned} E\Big(\sum_{\alpha\in\{s\}} p_\alpha y_\alpha\Big) &= \sum_{\alpha\in\{s\}} E(p_\alpha y_\alpha) \\ &= \frac{1}{n}\sum_{\alpha} w_\alpha y_\alpha \\ &= 1/\frac{1}{n}\sum_{\alpha} 1/y_\alpha \end{aligned}$$

Let $m(\{y_s\})$ denote this posterior expectation which is our estimate of the population mean. It follows from Jensen's inequality that $m(\{y_s\})$ must be smaller than the sample mean. Furthermore a standard calculation shows that the posterior variance is

$$v(\{y_s\}) = \frac{1}{n+1}\sum_{\alpha\in\{s\}} w_\alpha\big(y_\alpha - m(\{y_s\})\big)^2$$

Assuming that the posterior distribution is approximately normal this makes it easy to find an approximate 0.95 credible interval for either the population mean or population total.

To see how this estimator might work in practice we constructed a population by taking a random sample of size 500 from a log-normal distribution with mean and standard deviation of the log set equal to 5 and 0.6 respectively. Our estimator will be computed under the assumption that the sample was drawn using pps sampling with probabilities proportional to y. In addition to this design we considered two other situations where the selection probabilities were proportional to $y^{0.9}$ and $y^{1.1}$ respectively. For each of the three scenarios we took 1,000 random samples of size 25 and computed our point and interval estimates. The results are given in Table 5. Note we are assuming that the population size $N = 500$ is known. Our estimator works very well when the sampling design is pps proportional to y. But we see a bit of bias creeping in when the selection probabilities were pps proportional to y^c. This bias becomes worse as c moves farther away from 1 in either direction. In summary, our posterior will do a good job if it is roughly correct but can do a poor job when it is wrong. For the three simulations we computed the simple sample mean as well. Of course it is badly biased and its average absolute error was approximately 4.5 times as large our stepwise Bayesian estimator for this problem.

5 Discussion

Here we have argued that for some situations it makes sense to think of including the sampling design as part of the prior distribution. If in equation 3 we sum over π then the result is a prior distribution for y where order and the design no longer play a role. As we have seen, however, not margining out the design yields a convenient form for the posterior distribution. Furthermore, by considering stepwise Bayes models of the form of equation 3 we increase the types of prior information that can easily be incorporated into a Bayesian like analysis. These models are very flexible and avoid the difficult problem of specifying a single prior distribution. Given a sample it is often easy to simulate completed copies of the population from the resulting posteriors. This is a real strength of the Bayesian approach since using these simulated copies one can find approximately point and interval estimates of a variety of population parameters. Since this approach can sensibly express available prior information the resulting estimators will often have good frequentist properties.

References

Andreatta, G. and Kaufman, G. M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, 81:657–666.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A collection of problems from many fields for the student and research worker*. Springer-Verlag, New York.

Ghosh, J. K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by D. Basu*. Springer-Verlag, New York.

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

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

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18:379–390.

Pathak, P. K. (1961). Use of order-statistic in sampling without replacement. *Sankhyā A*, 23:409–414.

Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association*, 51:269–284.

Table 1: A comparison of the stB procedure with standard frequentist procedures for two different stratified designs. The results are based on 500 samples.

Est	Ave value	Ave aberr	Ave lowbd	Ave len	Freq of coverage
			4 strata of size 50		
strat	2,637	36.6	2,550	174	0.93
colspstr	2,637	36.6	2,579	115	0.78
reg	2,635	28.7	2,559	153	0.97
regstr	2,637	21.0	2,593	87	0.89
stB $a = 12$	2,637	21.4	2,585	104	0.95
			10 strata of size 20		
strat	2,636	21.5	2,580	111	0.94
colspstr	2,636	21.5	2,581	109	0.95
reg	2,636	19.6	2,558	156	1.00
regstr	2,635	24.6	2,599	72	0.74
stB $a = 12$	2,636	17.9	2,586	100	0.96

Table 2: A comparison of a stB procedure for five different choices of a with standard frequentist procedures. The results are based on 500 random samples of size 50.

Est	Ave value	Ave aberr	Ave lowbd	Ave len	Freq of coverage
usual	2,578	67.8	2,415	326	0.94
reg	2,574	30.8	2,498	152	0.94
			stB Model of equation 10		
$a = 4$	2,576	25.0	2,495	164	0.99
$a = 8$	2,575	17.3	2,518	114	0.98
$a = 12$	2,573	17.4	2,527	93	0.96
$a = 16$	2,573	18.0	2,532	82	0.92
$a = 20$	2,572	17.4	2,534	76	0.91

Table 3: A comparison of the stB procedure with standard frequentist procedures for the cubic population with population total equal to 1,143,957. the design selected 10 units at random from each of 4 strata of size 50. The results are based on 500 samples.

Est	Ave value	Ave aberr	Ave lowbd	Ave len	Freq of coverage
strat	1,163,664	21,757.68	1,112,834	101,661.4	0.94
colspstr	1,163664	21,757.68	1,137,833	51,662.09	0.64
reg	1,163,373	22,420.24	1,097,617	131,510.5	0.98
regstr	1,162,456	10,175.73	1,141,328	42,256.27	0.88
stB $a = 1.92$	1,163,073	9,442.34	1,135,955	54,393.89	0.96

Table 4: A comparison of the stB procedure with standard frequentist procedures for the sunspot population with population total equal to 21,510.2. The design selected 8 units at random from each of 6 strata of size 64. The results are based on 500 samples.

Est	Ave value	Ave aberr	Ave lowbd	Ave len	Freq of coverage
strat	21,409.0	1,471.21	17,798.76	7,220.49	0.94
colspstr	21,409.0	1,471.21	19,050.67	4,716.67	0.78
reg	21,397.4	1,480.61	17,557.52	7,679.80	0.95
regstr	21,752.8	1,264.15	19,405.11	4,695.36	0.85
stB $a = 0.25$	21,546.2	944.73	19,476.86	4,243.51	0.92

Table 5: The performance of the estimation procedure described in section 4.2 for 3 different sampling designs depending on y. The population total is 85,717. The results are based on 1000 random samples of size 25.

design	Ave value	Ave aberr	Ave lowbd	Ave len	Freq of coverage
$y^{0.9}$	83,607	8,883	61,668	43,879	0.923
y	86,474	9,145	63,672	45,604	0.958
$y^{1.1}$	89,678	9,805	65,857	47,641	0.945