Electronic Journal of Statistics
Vol. 12 (2018) 2515–2536
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1461

A noninformative Bayesian approach for selecting a good post-stratification

Patrick Zimmerman

Medtronic, Inc., Minneapolis, MN 55432, USA
e-mail: patrick.zimmerman@medtronic.com

and

Glen Meeden

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: gmeeden@umn.edu

Abstract: In the standard design approach to survey sampling prior information is often used to stratify the population of interest. A good choice of the strata can yield significant improvement in the resulting estimator. However, if there are several possible ways to stratify the population, it might not be clear which is best. Here we assume that before the sample is taken a limited number of possible stratifications have been defined. We will propose an objective Bayesian approach that allows one to consider these several different possible stratifications simultaneously. Given the sample the posterior distribution will assign more weight to the good stratifications and less to the others. Empirical results suggest that the resulting estimator will typically be almost as good as the estimator based on the best stratification and better than the estimator which does not use stratification. It will also have a sensible estimate of precision.

MSC 2010 subject classifications: Primary 62D05; secondary 62F15.
Keywords and phrases: Finite population sampling, stratification, prior information, stepwise Bayes.

Received March 2018.

1. Introduction

Stratification and post-stratification have a long history in survey sampling. Cochran (1977) discusses both approaches. Little (1993) develops a Bayesian model-based theory to be used with post-stratification. Elliott and Little (2000) and Si et al. (2015) give Bayesian approaches to modifying designed based sampling weights.

Often (post-)stratification is used for administrative convenience while in other cases auxiliary information makes it possible to divide a heterogenous population into strata which are internally homogenous with respect to a response variable of interest y. In the latter case there may be more than one plausible stratification available to the statistician, especially if several auxiliary variables are available with which to construct stratifications. For example, the population could be partitioned based on gender, age, or race categories. Various approaches can be used from relatively simple "binning" of population

units based on a single auxiliary variable to more sophisticated un-supervised dimension reduction techniques that can be used to generate a stratification based on several auxiliary variables (Pla (1991) used principal components analysis, and Golder and Yeomans (1973) used cluster analysis). However, none of these approaches are guaranteed to produce small within-stratum variability for y. All need to assume a model that relates y to the auxiliary variables used to form stratifications.

We are interested in the situation where the statistician has, prior to observing the sample, specified just a few possible partitions of the population. After one has observed the sample, say y_s, it is natural to want to use the observed y_s to select one of the partitions for making inferences. For example one might choose the stratification that minimizes the mean squared error within the strata. However, theory would be needed to justify this use of y_s and to derive an appropriate estimate of precision.

Here we present a method called "multiple post-stratification" which allows the statistician to simultaneously consider several stratifications that have been constructed before the sample was observed. Conceptually, multiple post-stratification is a simple idea: it uses a finite mixture model and statistical inference is done by averaging over the possible stratifications. We use a non-informative Bayesian approach in order to combine the finite mixture model concept with the objectivity of design-based estimators. We will show that the observed y_s can be useful in assessing the value of each possible stratification and we will show that the resulting estimator comes with a sensible estimate of precision. Although not post-stratification in the usual sense we believe it is a useful generalization of the concept.

We will begin by assuming that the design is simple random sampling and that there is just a small set of possible values for y and that they are known a priori. We will then show how our approach can be extended to other sampling designs and to a continuous y. This approach makes no model assumptions. It only assumes that a good stratification will have more homogenous strata (with respect to y) than a poor one.

In section 2 we introduce our multiple post-stratification approach. Our approach depends on the selection of a set of hyperparameters for our non-informative prior distributions. In section 3 we show how to choose these hyperparameters in an objective fashion. In section 4 we discuss how our estimator is computed. In section 5 we present some simulations that demonstrate how our approach could work in practice. Although our approach assumes that y takes on just a few possible values, in section 6, we show how it can be adapted to handle continuous variables and other designs. In section 7 we conclude with a few remarks.

2. Post-stratification, a Bayesian approach

Let $\mathcal{Y}$ denote the finite population of interest with labels $\mathcal{U} = \{1, 2, \ldots, N\}$ indexing the population, and with the response variable of interest denoted by

$y = (y_1, y_2, \ldots, y_N)$. In the exposition we will focus on the problem of estimating the population total, $T = \sum_{i=1}^{N} y_i$, under squared error loss given a random sample without replacement of size n. Later we will consider other possible sampling designs and note that our approach can be applied when estimating other population quantities of interest. We will write the set of sample indices, a subset of $\mathcal{U}$, as

$$s = \{i_1, i_2, \ldots, i_n\}$$

and the sampled part of y

$$y_s = \{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}$$

We denote the unsampled set of units, i.e. $\mathcal{U} \setminus s$, with s', and the unsampled part of y with $y_{s'}$.

Our primary interest is for situations where the possible values for a y_i is a categorical variable. We assume that $B = \{b_1, b_2, \ldots, b_r\}$, for some positive integer r, is the set of values which can be taken by y_i for any $i \in \mathcal{U}$. We will let $B = \{0, 1, \ldots, r-1\}$ for convenience, but an element $b \in B$ could be any real number. We also assume that r is small compared to n and that the values of B are known before the sample is taken. As will become clear in the following our method makes most sense when r is small. Formally this is a limitation of our approach but later we will show how it can be extended to situations where we have a continuous variable of interest.

Let $\mathcal{H}$ be a set of possible stratifications of y, and denote a generic element of $\mathcal{H}$ with h. That is, $h \in \mathcal{H}$ is a N-length vector where $h_i = j$ when the stratum membership of the i^{th} unit is j, for $i = 1, 2, \ldots, N$. We assume that the statistician has defined this set $\mathcal{H}$, possibly using auxiliary data that is known for the population, but is uncertain which will provide the greatest improvement in estimation precision. In addition, we assume that the number of possible stratifications belonging to $\mathcal{H}$ is small. For simplicity, we will assume for now that each stratification has the same number of strata, say k. For any set A, stratification $h \in \mathcal{H}$, and stratum $j = 1, 2, \ldots, k$, we define A_{jh} as the subset of A that lies in stratum j according to h. This use of notation allows us to easily refer to subsets like $\mathcal{U}_{jh}$, s_{jh}, and s'_{jh}, for example. We also use this subscript pair on N_{jh} and n_{jh} to denote the population and sample sizes associated with stratum j as defined by h, respectively. We are assuming that all of the N_{jh}'s are known.

Since we will be taking a Bayesian approach we need to specify a prior distribution for our problem. We will do this in two stages. First we define a prior over $\mathcal{H}$. At the second stage we need to define the conditional distribution for y given $h \in \mathcal{H}$. More formally we write

$$\begin{aligned} \Pr(y) &= \sum_{h \in \mathcal{H}} \Pr(h) \Pr(y|h) \\ &= \sum_{h \in \mathcal{H}} \Pr(h) \prod_{j=1}^{k_h} \Pr(y_{\mathcal{U}_{jh}}|h) \end{aligned}$$

The first line above represents the finite mixture model approach; each possible stratification in $\mathcal{H}$ contributes a model. The second line above represents (in a Bayesian manner) the idea behind stratification: given $h \in \mathcal{H}$, y can be split into the independent strata $y_{\mathcal{U}_{1h}}, y_{\mathcal{U}_{2h}}, \ldots, y_{\mathcal{U}_{k_h h}}$.

Although, in some cases, a statistician may be comfortable defining the prior $\Pr(h)$ across $\mathcal{H}$ based on subjective belief or past performance of the stratifications we acknowledge that this will not always be true. As a default, we will assume the uniform prior distribution over $\mathcal{H}$ in keeping with our non-informative Bayesian approach. Keep in mind that $\mathcal{H}$ should be a smallish set of possible stratifications based on the auxiliary information at hand.

To complete the definition of $\Pr(y)$, we need to define $\Pr(y_{\mathcal{U}_{jh}}|h)$ for each $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$. For ease of exposition we will restrict ourselves to the case when $r = 2$. For this case we use the Beta-Binomial model.

$$\begin{aligned} \theta_{jh} &\sim \text{Beta}(\epsilon_{jh}, \epsilon_{jh}) \text{ independently for } h \in \mathcal{H}, j = 1, 2, \ldots, k \\ y_i|\theta_{jh} &\sim \text{Bernoulli}(\theta_{jh}) \text{ independently for } i \in \mathcal{U}_{jh} \end{aligned}$$

where ϵ_{jh} for $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$ are known hyperparameters. Note each stratification has k of these hyperparameters; one for each stratum. In section 3 we will discuss how they should be chosen so that the resulting estimators will have good frequentist properties.

We can integrate across θ_{jh} to obtain a concise expression of $\Pr(y_{\mathcal{U}_{jh}}|h)$. For a given vector y, a set of units $A \subseteq \mathcal{U}$, and a real number b, let $c_y(b, A)$ denote the number of units i in a set $A \subseteq \mathcal{U}$ where $y_i = b$. Then,

$$\Pr(y_{\mathcal{U}_{jh}}|h) = \frac{\Gamma(2\epsilon_{jh})\Gamma(\epsilon_{jh} + c_y(1, \mathcal{U}_{jh}))\Gamma(\epsilon_{jh} + c_y(0, \mathcal{U}_{jh}))}{\Gamma(\epsilon_{jh})^2\Gamma(2\epsilon_{jh} + N_{jh})} \tag{1}$$

Standard calculations yield a similar formula for the sample

$$\Pr(y_{s_{jh}}|h) = \frac{\Gamma(2\epsilon_{jh})\Gamma(\epsilon_{jh} + c_y(1, s_{jh}))\Gamma(\epsilon_{jh} + c_y(0, s_{jh}))}{\Gamma(\epsilon_{jh})^2\Gamma(2\epsilon_{jh} + n_{jh})} \tag{2}$$

Now, we can discuss the posterior distribution $\Pr(y_{s'}|y_s)$. Like the marginal prior distribution of y_s, $\Pr(y_{s'}|y_s)$ also maintains the general structure of the prior distribution. That is,

$$\begin{aligned} \Pr(y_{s'}|y_s) &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \Pr(y_{s'}|y_s, h) \\ &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \frac{\Pr(y|h)}{\Pr(y_s|h)} \\ &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \frac{\prod_{j=1}^{k_h} \Pr(y_{\mathcal{U}_{jh}}|h)}{\prod_{j=1}^{k_h} \Pr(y_{s_{jh}}|h)} \\ &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \prod_{j=1}^{k_h} \Pr(y_{s'_{jh}}|y_{s_{jh}}, h) \end{aligned}$$

We see that it is a finite mixture model where each possible stratification $h \in \mathcal{H}$ supplies a different model for the unseen units, i.e. $y_{s'}|y_s, h$. Secondly, the unseen units from different strata are conditionally independent given some $h \in \mathcal{H}$.

In the posterior distribution $y_{s'}|y_s$, the probability $\Pr(h|y_s)$ for some $h \in \mathcal{H}$ can be thought of as the mixture weight for h in a finite mixture model. This probability is proportional to $\Pr(h)\Pr(y_s|h)$ where $\Pr(h)$ is a known prior distribution, so we can see that $\Pr(y_s|h)$ is how the observed data help determine the mixture weights. That is

$$\Pr(h|y_s) \propto \Pr(h) \prod_{j=1}^{k_h} \frac{\Gamma(2\epsilon_{jh})\Gamma(\epsilon_{jh} + c_y(1, s_{jh}))\Gamma(\epsilon_{jh} + c_y(0, s_{jh}))}{\Gamma(\epsilon_{jh})^2\Gamma(2\epsilon_{jh} + n_{jh})} \tag{3}$$

Note that $\Pr(y_s|h)$ will be large when the composition of y_s within the strata defined by h is relatively homogenous (when $\bar{y}_{s_{jh}}$ is close to zero or one) compared to the composition for other stratifications. So, stratifications that separate y_s into "homogenous groups" will have relatively large mixture weights compared to those which do not. $\Pr(y_s|h)$ will also depend on the relationship between the sample allocation of y_s with respect to h and the choice of hyperparameters, and we will discuss this in section 3.

Using Equations (1) and (2) we find that

$$\Pr(y_{s'_{jh}}|y_{s_{jh}}, h) = \frac{\Gamma(2\epsilon_{jh} + n_{jh})\Gamma(\epsilon_{jh} + c_y(1, \mathcal{U}_{jh}))\Gamma(\epsilon_{jh} + c_y(0, \mathcal{U}_{jh}))}{\Gamma(2\epsilon_{jh} + N_{jh})\Gamma(\epsilon_{jh} + c_y(1, s_{jh}))\Gamma(\epsilon_{jh} + c_y(0, s_{jh}))}$$

In the same way, we can easily calculate the conditional posterior distribution of a single unseen unit. For $h \in \mathcal{H}$, $i \in s'_{jh}$, and $z \in \{0, 1\}$,

$$\begin{aligned}\Pr(y_i = z|y_s, h) &= \frac{\Gamma(2\epsilon_{jh} + n_{jh})\Gamma(\epsilon_{jh} + c_y(z, s_{jh}) + 1)}{\Gamma(2\epsilon_{jh} + n_{jh} + 1)\Gamma(\epsilon_{jh} + c_y(z, s_{jh}))} \\ &= \frac{\epsilon_{jh} + c_y(z, s_{j,h})}{2\epsilon_{jh} + n_{jh}}\end{aligned}$$

and from this one can find the posterior expectation of an unsampled unit.

The posterior distribution $y_{s'}|y_s$ can be used to estimate any parameter $\gamma(y)$ under a variety of loss functions, but we will just consider the squared-error loss function (which implies that the Bayes rules will be posterior expectations). Because every possible y_s has positive probability under our prior distribution, the Bayes rule under squared-error loss will be unique, and hence admissible. For estimating the population total, say $T(y)$, the Bayes rule is

$$\begin{aligned}\mathrm{E}[T(y)|y_s] &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \sum_{i \in \mathcal{U}} \mathrm{E}[y_i|y_s, h] \\ &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \sum_{j=1}^{k_h} \left(n_{jh}\bar{y}_{s_{jh}} + (N_{jh} - n_{jh}) \frac{\epsilon_{jh} + c_y(1, s_{j,h})}{2\epsilon_{jh} + n_{jh}} \right) \\ &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \sum_{j=1}^{k_h} N_{jh} \left(\frac{n_{jh}}{N_{jh}} \bar{y}_{s_{jh}} + \frac{N_{jh} - n_{jh}}{N_{jh}} \frac{\epsilon_{jh} + c_y(1, s_{j,h})}{2\epsilon_{jh} + n_{jh}} \right)\end{aligned}$$

As $\epsilon_{jh} \to 0$ for each $j = 1, 2, \ldots, k$ and $h \in \mathcal{H}$, this estimator will converge to a weighted average of design-based stratified estimators of μ where the weights are $\Pr(h|y_s)$.

For the general $r > 2$ the story is very similar except that we will be using a Dirichlet distribution of the appropriate dimension rather than the beta distribution. For any particular stratum within any particular stratification the values of the units in the parameter vector defining the Dirichlet distribution will be constant.

The model we've described here works fine when the set B of possible values is known ahead of time. That is, we assume that each element of y takes a value from the set B, and then the posterior distribution will give positive probability to each $b \in B$. Note, however, that there is positive probability that not all the possible values for y will appear in every sample. When this happens, as we have seen, our Bayes model gives positive posterior probability to all the values of B. If we are trying to be "objective" and mimic as far as possible the standard frequentist estimator this should not happen. Our posterior distribution should only give positive probability to the values of B which actually appeared in the sample. This means that estimators based on such a posterior distribution will not be a Bayes rule for any given prior distribution.

In order to justify such a posterior and the resulting estimator we need to use the stepwise Bayes approach. Johnson (1971) presented an early special case of this technique for use when estimating the mean of a Binomial random variable, and Hsuan (1979) explained the stepwise Bayes idea in a more general decision theory context. It has been used in survey sampling problems to prove the admissibility of many of the standard frequentist estimators. Ghosh and Meeden (1997) give several examples and show how it is related to the Bayesian bootstrap of Rubin (1981). An early example of the type of parameter spaces used in these proofs can be found in Hartley and Rao (1968). A proof that the posterior defined in Section 2 has a stepwise Bayes justification and provides admissible estimators is given in Zimmerman (2013). We should note however that the admissibility proof does not apply for our estimator in section 6 when y is a continuous variable.

3. Choosing hyperparameters

In the models presented above, a set of hyperparameters $\epsilon_{1h}, \epsilon_{2h}, \ldots, \epsilon_{kh}$ is associated with each stratification $h \in \mathcal{H}$. We will now argue that there is a non-informative way to select them. Our recommended choice of hyperparameters is based on separately considering how the choice affects the distributions of $y|(y_s, h)$ and $h|y_s$.

First we consider the impact of the choice of ϵ_{jh} on $y|(y_s, h)$. If we examine the posterior probability $\Pr(y_{s'_{jh}}|y_{s_{jh}}, h)$ discussed in the previous section, we can see that choosing ϵ_{jh} to be small for $j = 1, 2, \ldots, k$ will make inference, conditional on a given h, agree with design-based (post)-stratification using that same h. The relative sizes of $\epsilon_{1h}, \epsilon_{2h}, \ldots, \epsilon_{kh}$ is not important in this regard, so long as they are sufficiently small.

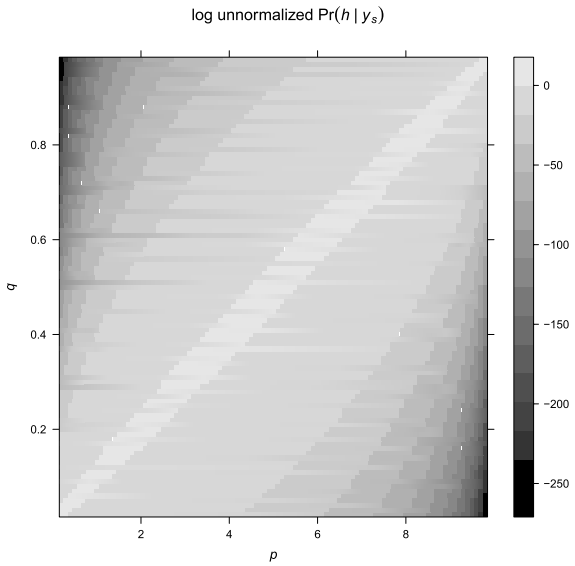

FIG 1. *Log-unnormalized posterior probability of a two-stratum stratification h given population and sample proportions of units falling in its first stratum (p and q, respectively).*

Now, we consider the relationship between hyperparameter choice and the distribution $h|y_s$. This relationship is more complicated. Reviewing Equation (3), we can see that the factors influencing the behavior of $\Pr(h|y_s)$ can be separated into three categories: the prior distribution $\Pr(h)$, the within-strata homogeneity of $y_s|h$, and the sample allocation of $y_s|h$ (i.e. the number of units in each sample stratum as defined by h). The prior distribution $\Pr(h)$ is not related to choice of ϵ_{jh}. The relationship between the within-strata homogeneity of $y_s|h$ and $\Pr(h|y_s)$ is sensible: as homogeneity increases, $\Pr(h|y_s)$ increases. Although hyperparameter choice may affect the degree to which $\Pr(h|y_s)$ rewards within-strata homogeneity, the preference for homogeneity will always exist. Finally, hyperparameter choice strongly affects the relationship between the sample allocation of $y_s|h$ and $\Pr(h|y_s)$. In what follows, we show that choosing ϵ_{jh} to be proportional to N_{jh} will make the distribution $h|y_s$ reward stratifications for which the sample allocation of $y_s|h$ is close to proportional. Given these considerations, we recommend setting $\epsilon_{jh} = \epsilon N_{jh}/N$ for each $h \in H$ and $j = 1, 2, \ldots, k_h$ where ϵ is small. We will now discuss in some detail how to select a value of ϵ.

First, we present an example to help orient the reader for the theoretical analysis that will follow. Suppose that $k = 2$, that $n = 100$, that $\epsilon_{jh} = N_{jh}/N$ for some $h \in H$, and that y_s consists of all distinct values so that with-stratum homogeneity plays no role.

In this case, if we imagine varying h to achieve a variety of associated population and sample allocations, $\Pr(h|y_s)$ can be thought of as a function of $p = N_{1h}/N$ and $q = n_{1h}/n$. Figure 1 presents this graphically. We can see that

the posterior probability of h is large when the sample allocation is close to proportional allocation and small when it is not. More specifically, two properties are evident. First, for a fixed p, the posterior probability appears to be a convex function of q that achieves its maximum at $q = p$. Second, letting p vary, the posterior probability appears to be constant along the line defined by $p = q$. It turns out that both of these properties hold for any pair of k and n when n is sufficiently large (they correspond to results (i) and (ii) in the Theorem below).

Now, before stating our theorem, we will take a moment to to review how asymptotics is usually done in a survey sampling context. We will define a sequence of finite populations, each one associated with a response, a set of stratifications, and a sample. The notation and general set-up we use here is similar to that employed by Fuller (2009) when working with survey sampling asymptotics. Let $y^{(N)}$ be a vector containing the first N terms from the infinite sequence $y^{(\infty)} = \{y_i\}_{i=1}^{\infty}$ for $N = 1, 2, \ldots$. Also, let $\mathcal{H}^{(\infty)}$ be a finite set of infinite sequences that stratify $y^{(\infty)}$ into k strata, and let $H^{(N)}$ be the set that contains, for each sequence $h^{(\infty)} \in \mathcal{H}^{(\infty)}$, a vector $h^{(N)}$ of the first N terms from $h^{(\infty)}$. So, each $N \in \{1, 2, \ldots\}$ is associated with a finite population response and a set of stratifications. Note that one or more of the k strata defined by $h^{(\infty)} \in \mathcal{H}^{(\infty)}$ will not appear in $h^{(N)}$ for small enough values of N, but we ignore this problem because, with only finitely stratifications in $\mathcal{H}^{(\infty)}$, we can find an N_0 large enough so that all k strata appear in each $h^{(N)} \in \mathcal{H}^{(N)}$ for all $N \geq N_0$. Next, for $h^{(N)} \in \mathcal{H}^{(N)}$, let $p_{jh}^{(N)} = N_{jh}/N$, i.e. the proportion of units from the N^{th} population falling in the j^{th} stratum defined by $h^{(N)} \in \mathcal{H}^{(N)}$, for $j = 1, 2, \ldots, k$. Let the hyperparameter for the N^{th} population, $\epsilon_{jh}^{(N)}$, be set equal to $\epsilon p_{jh}^{(N)}$, for some $\epsilon > 0$, so that it is proportional to the population stratum sizes. Then, for some fixed $f \in (0, 1)$, let $s^{(N)}$ be a sample of size $n^{(N)} = [fN]$, i.e. the largest integer less than fN, and set $q_{jh}^{(N)} = n_{jh}^{(N)}/n^{(N)}$ where $n_{jh}^{(N)}$ is the number of sample units that fall in the j^{th} stratum defined by $h^{(N)} \in \mathcal{H}^{(N)}$, for $j = 1, 2, \ldots, k$. Although the stratum sample sizes for a variety of index pairs jh will be less than two for small enough N, we ignore this problem, too, because it will not be an issue once N reaches some finite threshold. Finally, write $p_h^{(N)} = (p_{1h}^{(N)}, p_{2h}^{(N)}, \ldots, p_{kh}^{(N)})$ and $q_h^{(N)} = (q_{1h}^{(N)}, q_{2h}^{(N)}, \ldots, q_{kh}^{(N)})$.

Let $\mathcal{S}_{(k-1)}$ be the $(k-1)$-dimensional unit simplex, let $\mathcal{S}_{(k-1)}^0$ denote its interior. With this bit of additional notation we now state the theorem. Note the proof depends on the lemma given in Appendix A.

Theorem. *Assume that, for each $h^{(\infty)} \in \mathcal{H}^{(\infty)}$, there is some fixed $p_h, q_h \in \mathcal{S}_{(k-1)}^0$ such that $p_h^{(N)} \to p_h$ and $q_h^{(N)} \to q_h$ as $N \to \infty$. Also, assume that $y^{(\infty)}$ consists of distinct values, and that $\Pr(h^{(N)})$ is uniform across $H^{(N)}$ for each $N = 1, 2, \ldots$.*

Pick an arbitrarily small $\eta > 0$. Then, for a sufficiently large N_0, the following two properties hold for all $N > N_0$:

(i) Suppose that, for some pair $h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}$ and some $\lambda \in (0, 1]^k$, $p_h = p_{h'}$ and $q_{jh'} = q_{jh} + \lambda_j(p_{jh} - q_{jh})$ for $j = 1, 2, \ldots, k$. In other

words, the limiting population stratum allocation is the same for h and h', and the limiting sample allocation for h' either lies between that of h and proportional allocation or is equal to proportional allocation. Then, $\Pr(h'^{(N)}|y_{s^{(N)}}) > \Pr(h^{(N)}|y_{s^{(N)}})$.

(ii) If, for some pair $h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}$, $p_h = q_h$ *and* $p_{h'} = q_{h'}$, *then,* $|\Pr(h^{(N)}|y_{s^{(N)}}) - \Pr(h'^{(N)}|y_{s^{(N)}})| < \eta$.

Proof. First, note that the set-up and assumptions above imply that, for $N = 1, 2, \ldots,$

$$\begin{aligned}
\frac{\log \Pr(h^{(N)}|y_{s^{(N)}})}{n^{(N)}} &\propto \frac{1}{n^{(N)}} \sum_{j=1}^{k} \log \\
&\quad \left(\frac{\Gamma(n^{(N)}\epsilon p_{jh}^{(N)})\Gamma(\epsilon p_{jh}^{(N)} + 1)^{n_{jh}^{(N)}} \Gamma(\epsilon p_{jh}^{(N)})^{(n^{(N)} - n_{jh}^{(N)})}}{\Gamma(\epsilon p_{jh}^{(N)})^{n^{(N)}} \Gamma(n^{(N)}\epsilon p_{jh}^{(N)} + n_{jh}^{(N)})} \right) \\
&\propto \frac{1}{n^{(N)}} \sum_{j=1}^{k} \log \left(\frac{\Gamma(n^{(N)}\epsilon p_{jh}^{(N)})(\epsilon p_{jh}^{(N)})^{n_{jh}^{(N)}}}{\Gamma(n^{(N)}\epsilon p_{jh}^{(N)} + n_{jh}^{(N)})} \right) \\
&\propto \sum_{j=1}^{k} q_{jh}^{(N)} \log(\epsilon p_{jh}^{(N)}) + \frac{\log \Gamma(n^{(N)}\epsilon p_{jh}^{(N)})}{n^{(N)}} \\
&\qquad - \frac{\log \Gamma(n^{(N)}(\epsilon p_{jh}^{(N)} + q_{jh}^{(N)}))}{n^{(N)}} \\
&\propto f_{n^{(N)}}(p_h^{(N)}, q_h^{(N)})
\end{aligned}$$

where f_m is the function from the lemma if we choose the ϵ shown here to equal the v from the lemma. So, for any pair $h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}$ and any $N \in \{1, 2, \ldots\}$,

$$\begin{aligned}
&\log \Pr(h^{(N)}|y_{s^{(N)}}) - \log \Pr(h'^{(N)}|y_{s^{(N)}}) \\
&\quad \propto n^{(N)}(f_{n^{(N)}}(p_h^{(N)}, q_h^{(N)}) - f_{n^{(N)}}(p_{h'}^{(N)}, q_{h'}^{(N)}))
\end{aligned}$$

Now, let δ be the minimum non-zero value of $|f(p_h, q_h) - f(p_{h'}, q_{h'})|$ for any pair $h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}$. Then, since f_m is continuous and converges uniformly on a set containing $\{(p_h, q_h) : h^{(\infty)} \in \mathcal{H}^{(\infty)}\}$, we can find N_1 such that, for $N > N_1$ and any pair $h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}$,

$$|f_{n^{(N)}}(p_h^{(N)}, q_h^{(N)}) - f_{n^{(N)}}(p_{h'}^{(N)}, q_{h'}^{(N)}) - (f(p_h, q_h) - f(p_{h'}, q_{h'}))| < \delta$$

Now, if $h^{(\infty)}$ and $h'^{(\infty)}$ have (p_h, q_h) and $(p_{h'}, q_{h'})$ that fit the scenario described for result (i), our lemma proves that $f(p_h, q_h) < f(p_{h'}, q_{h'})$, and our choice of N_1 implies that $f_{n^{(N)}}(p_h^{(N)}, q_h^{(N)}) < f_{n^{(N)}}(p_{h'}^{(N)}, q_{h'}^{(N)})$ for $N > N_1$. Hence, $\Pr(h^{(N)}|y_{s^{(N)}}) < \Pr(h'^{(N)}|y_{s^{(N)}})$ for $N > N_1$.

Next, for any pair $h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}$ that fit the scenario described for result (ii),

$$\log \Pr(h^{(N)}|y_{s^{(N)}}) - \log \Pr(h'^{(N)}|y_{s^{(N)}}) \propto n^{(N)} d_{n^{(N)}}(p_h^{(N)}, p_{h'}^{(N)})$$

where d_m is the function from the lemma. Note that d_m is continuous and uniformly has magnitude $o(1/m)$ on a set containing $\{(p_h, p_{h'} : h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}\}$. Hence, we can find N_2 such that, for any $h^{(\infty)}, h'^{(\infty)} \in \mathcal{H}^{(\infty)}$ that fit the scenario described in result (ii) and $N > N_2$,

$$\log \Pr(h^{(N)}|y_{s^{(N)}}) - \log \Pr(h'^{(N)}|y_{s^{(N)}}) < \eta$$

Finally, we can simply set $N_0 = \max(N_1, N_2)$, so that both results (i) and (ii) hold for $N > N_0$. □

This theorem essentially shows that the properties of $\Pr(h|y_s)$ evident in Figure 1 hold approximately for any k when n is large (the convergence of $\Pr(h|y_s)$ to its limiting form only depends on the size of n; we only dealt with an increasing N because $N > n$ must be true). Therefore, our recommendation to define $\epsilon_{jh} = \epsilon N_{jh}/N$ for a small ϵ and $j = 1, 2, \ldots, k$ for each $h \in \mathcal{H}$, achieves desirable behavior from both $\Pr(y|y_s, h)$ and $\Pr(h|y_s)$.

But what is a good choice for a small value of ϵ? Consider the again equation (3) and the term $\Gamma(2\epsilon_{jh}/\Gamma(\epsilon_{jh})^2$. This is an increasing function of ϵ_{jh} whose limit is zero as ϵ_{jh} approaches zero. It takes on the value one when $\epsilon_{jh} = 1$ and increases rapidly from that point. Since a stratification with k strata has k such terms choosing a small value of ϵ will typically increase the posterior probability given to stratifications with fewer strata. If all the stratifications have about the same number of strata this will not matter much but if the numbers of strata in the stratifications under consideration vary widely it is not clear how to select a good choice of ϵ. In his discussion of stratification in Section 5A.8 (Cochran, 1977) Cochran indicates that, typically, most of the precision gain from stratification is obtained with six or fewer strata. We believe that this is will often be true when a statistician is considering a small number of possible post-stratifications. In the simulations that follow we never consider stratifications with more than 8 strata. For such situations we have found that choosing any value of ϵ between 0.1 and 1 works well and in all of our examples the results are very robust against the particular choice in this interval.

4. Computing the estimator

Here we briefly discuss how our estimator can be computed. Let H be the number of possible stratifications belonging to $\mathcal{H}$. Given a sample, let λ_h be the the posterior probability our model assigns to stratum h for $h = 1, \ldots, H$. Then $\lambda = (\lambda_1, \ldots, \lambda_H)$ is the posterior distribution over $\mathcal{H}$ given the sample. This is easily computed under our model. When estimating an arbitrary $\gamma(y)$ then one may proceed in the usual Bayesian fashion. First one selects a stratification using the distribution λ. Given the stratification one then uses our posterior distribution in each stratum to generate a complete set of possible values. One then finds $\gamma(y)$ for this simulated complete copy of the population. One repeats this many times to get a large set of simulated values for $\gamma(y)$. The mean of this set will be our point estimate and an approximate 0.95 Bayesian credible

interval is given by the interval defined by the lower 0.025 quantile and the upper 0.975 quantile of this set.

When one is estimating the population total, $T(y)$, there is a quicker way to get our estimates. Under our model, for a given stratification h, we can calculate directly the posterior expectation of the population total and its posterior variance, say t_h and vr_h. Then our point estimate of the population total is just

$$t = \sum_{h=1}^{H} \lambda_h t_h$$

To get the posterior variance of t we use the well known fact that a variance is equal to the sum of the variance of a conditional expectation plus the expectation of the conditional variance. So conditioning on a possible stratification we find that

$$vr = \sum_{h=1}^{H} \lambda_h (t - t_h)^2 + \sum_{h=1}^{H} \lambda_h vr_h$$

When computing our interval estimate we will assume that t is approximately normally distributed with variance vr and use the standard design based formula to get our interval. We have seen that this yields estimates which closely approximate the more complete Bayesian analysis and they can be found more quickly.

The only non-routine part of this calculation is finding λ the vector of posterior probabilities. Some *R* code (R Core Team, 2017) that does this is given in Appendix B.

5. Simulations

To see how our approach could work in practice we used the data set *nhanes* which is available in R. This data set is discussed in Lumley (2010) and is a subset of a much larger set which was the result of a cluster sample. We will consider this set as our population and use it in our simulation studies. We will not make any model assumptions about the variables in the population.

The set contained information on 8,591 individuals for seven different variables. For each individual we have their gender (1 = male and 2 = female), age category in years ((0, 19], (19, 39], (39, 59] and (59, Inf]), and race (Hispanic, non-Hispanic white, non-Hispanic black and other). In addition we know their designated primary sampling status within a household cluster. Most of these were either 1 or 2 with a few 3's. We combined the 3's with the 1's to get the stratification, psu.

The y variable of interest was 1 if the total cholesterol of an individual was over 240 mg/dl and 0 otherwise. We denote this variable, by HiC. Some of the individuals had missing values primarily in the HiC category. After removing these individuals there remained a population of size 7,846.

Stratifying on gender gives us two strata of sizes 3889 and 3957. We denote this stratification by g. Stratifying on race gives us four strata of sizes 2532, 3450,

1406 and 458. We combined strata 1 with 3 and 2 with 4 to get a stratification on race with just two strata. We denote these stratifications by r4 and r2. Stratifying on age yields four strata of sizes 2150, 1905, 1911 and 1880. Again we constructed a two set stratification by combing the sets 1 and 2 and by combining sets 3 and 4. These two stratifications are denoted by a4 and a2. The sizes of the two strata in the stratification based on psu is 3906 and 3940. This gives us five stratifications, three with two strata and two with three strata. We constructed two more stratifications having eight strata. In the first we crossed race (with two levels), age (with two levels) and gender. To get the second we replaced race with psu. These stratifications are denoted by r2 $\times$ a2 $\times$ g and psu $\times$ r2 $\times$ g.

We now have seven possible stratifications which we will order as follows: psu $\times$ r2 $\times$ g, r2 $\times$ a2 $\times$ g, r4, r2, a4, a2 and g. The correlation of these stratifications with the y value HiC are 0.03, 0.13, -0.02, 0.03, 0.19, 0.2 and 0.02. The only stratifications that seem to contain some information about HiC are those that contain the age variable.

We took 500 simple random samples of size 400 from our population. The true population total of HiC is 787 and the average absolute error of the standard estimator, population size times the sample mean, in our simulations was 91.2. The average absolute error using the a4 stratification was 88.4 while for the a2 stratification it was 89.4. For each sample we calculated the posterior probability of each stratification under our model. The average posterior probabilities, over the 500 samples, given to stratifications a4 and a2 were 0.500 and 0.496 respectively. The average absolute error of our method was 89.1. Another thing one could do is just select the stratification with the largest posterior probability. In the 500 samples our model selected stratification a4 247 times and selected a2 502 times. The average absolute error for this procedure was 88.9. Even though there is not a lot of information about HiC in the stratifications our method preforms almost as well as knowing the best stratification. We will not present the frequency of coverage for our intervals but they covered about 95% of the time because they are essentially based on standard frequency theory.

The basic assumption underlying our approach is that within each stratum, from any stratification, the y values are roughly exchangeable. That is why we have focused on simple random sampling as the design. Our approach can handle missing observations as long as one is willing to assume that observations are missing at random within each stratum. Note that given a sample our posterior distribution over the stratifications does not explicitly depend on the design. Implicitly it does of course, because the design can affect what y values appear in a stratum. So our approach could be used when the design weights for the units in the sample are unknown. To see what might happen we repeated the simulation with two different sampling plans. In the first, individuals in the second stratum of r2 was twice as likely to be selected as individuals in the first stratum. In the second, individuals in the second stratum of a2 was twice as likely to be selected as individuals in the first stratum. In each stratification when computing the estimate we did not use the sampling weights. The results

were very similar to those above in that our method did almost as well as knowing the best stratification.

For a second example we used another variable in the data set *nhanes* called sampling weights, say smpwt, which comes from the design. This is a continuous variable which we converted to a 0–1 variable by setting everything above the median equal to 1 and the rest to 0. We denote this new variable ysmpwt. The correlations of this variable with our seven stratifications are 0.32, 0.2, 0.23, 0.53, 0.11, 0.02 and 0.02. So in this case race seems to contain the most information about the y variable. We repeated our simulations under our three sampling plans. Under simple random sampling the average absolute error of the sample mean was 152.9. The average absolute error under the stratifications r4 and r2 were 128.0 and 130.0 respectively. The average posterior probability for r4 and r2 was 0.23 and 0.77 respectively. The average absolute error for our procedure was 137.4. So again we are doing almost as well as knowing the best stratification. Here the improvement over simple random sampling is a bit more because there is more information about y in the stratification. The results for the other two sampling plans were very similar in that we do almost as well as knowing the best stratification.

In the results of the previous paragraph we have an example where one sees that our approach tends to prefer stratifications with fewer sets over those with more sets. One can dampen this effect by using an $\epsilon > 1$. A good choice of an ϵ will depend on the range of the numbers of strata in the possible stratifications under consideration and a poor choice can lead to poor results when stratifications with more strata are not as good as ones with with fewer strata. This issue needs further study and this is why we have only considered situations where the set of possible stratifications each contain about the same number of strata.

6. Continuous populations and other designs

6.1. Continuous populations

Earlier, we stated that our method would not work when y was a continuous variable. Technically this is correct because if all the y values are distinct there is no "*clumping*" of the y values which allows us to identify the better stratifications. This is because the posterior distribution only discerns between equal and unequal values; there is no measurement of how unequal two values are. One way to overcome this difficulty is to map the set of distinct values in y_s to a smaller set of "bins" (i.e. intervals). We then use this set of possible "bins" to compute our posterior distribution over the possible stratifications. But of course we would use the actual observed values in y_s values when computing the estimator for a given stratification.

There are many ways that one could discretize the sample. Here is one simple approach that seems to work well. Suppose for our sample of size n we want to replace the observed sample values with r values where $n = \nu \times r$ and ν and r are integers. One way do this is to take the ν smallest values in the sample and

assign them the value 1. We then take the next ν smallest values and assign them the value 2. We continue in this way until the ν largest values are assigned the value r.

At first glance this might seem somewhat arbitrary but this is not really the case. When calculating the posterior distribution, what is important is the frequency of each discretize value in the sample. The actual values play no role in the calculation. Given the sample size what is important is not to pick an r that is too small or too large. If it is too small then you will be throwing away too much information. On the other hand if it is too large then there will not be enough clumping in the y_s to pick out the better stratifications.

As an example of a continuous variable for y we took the variable smptwt/10,000 and used the same set of 7 possible stratifications. The correlations between this variable and our seven stratifications were 0.37, 0.25, 0.20, 0.61, 0.09, 0.08 and 0.02. Here we see that race seems to contain the most information about y.

One thing that we need to decide is how many groups we should use. This will depend in part on the sample size and the nature of the population. For a sample of size 400 we first considered the case with 10 groups of size 40. We repeated our simulation generating 500 simple random samples. The true population total is 25,535 and the average absolute error of the standard estimator of the population total was 738.2. The average absolute errors for the first four stratifications were 637.4, 638.3, 618.6, and 626.6. Absolute errors were considerably larger for the remaining three stratifications. Our method selected the stratification r2, the best one, in every sample. We repeated the simulation where the number of groups were 40. The average absolute error for the various stratifications was very similar to the previous case. Here stratification r2 had the largest posterior probability in 458 of the 500 samples. Stratifications r4 and a4 had the largest posterior probabilities 39 and 3 times respectively. We also repeated the simulations where the number of groups were 2, 5 and 20. The results were very similar to those just above: in all cases our method did almost as well as knowing the best stratification. So our approach appears to be fairly robust against the choice of the number of groups and in most cases there should be a reasonably wide range of choices that are close to the best.

A possible concern for our approach is what happens when all the possible stratifications are equally good or equally bad. Although we will not present any simulations to show this, what happens is what you would expect. The posterior distribution over the possible stratifications tends to be uniform and there is no gain in efficiency; there is just the extra cost of computing the posterior distribution.

6.2. Other designs

Our approach makes no explicit use of the design probabilities or related weights. It only uses the amount of homogeneity within strata. So far we have focused on simple random sampling because it ensures approximate proportional allocation.

Although this is a good property it is not necessary for our method to work well. To see what might happen under other designs we return to the y variable defined just above which is equal to smptwt/10,000. For the set of possible stratifications we took the set of 6 stratifications formed by crossing all possible pairs of psu, r2, a2 and g. The correlations between y and these six stratifications, psu $\times$ r2, psu $\times$ a2, psu $\times$ g, r2 $\times$ a2, r2 $\times$ g and a2 $\times$ g were 0.34, 0.1, 0.08, 0.56, 0.56 and 0.08. Note that the three stratifications that include r2 have the largest correlations and they are the best stratifications. In 500 simple random sample of size 400 the ratio of each of their average absolute errors for estimating the population total to one of the other three average absolute errors was about 0.75.

For each of our four stratifications with two sets (psu, a2, r2 and g) we considered the design where the inclusion probability for units in the second stratum was twice that of those in the first. We also considered two designs which depended on y. Let y = smptwt/10,000. The range of possible values for y is from 0.43 to 15.81. In the first design the probability that y_i appeared in the sample was proportional to $y_i + 14$ while in the second it was $30 - y_i$. So in the first case we will oversample units with larger values of y_i while in the second the units with small value of y_i will be oversampled. For each of these designs we took 500 random samples of size 400. In all cases our posterior distribution put all of its probability on the three stratifications that included race. The stratification r2 $\times$ g received about two thirds of the probability and the other two containing race each about one sixth. So our approach had no difficulty identifying the good stratifications.

As another test of our approach we repeated the above simulations with the same sampling designs but with just two possible stratifications. The first was r2 $\times$ g and the second used a stratification with four equal sized strata based on the ordered values of y. The correlation of this second stratification with y was 0.85. Clearly this second stratification is the better one and in our simulations for every design and for every sample our approach gave it a posterior probability of one, when the calculation was done to three significant figures.

6.3. Discussion

Stratification is a classic method for improving estimates in finite population sampling. Often the stratification of interest is constructed using the information in an auxiliary variable, say x. We have been interested in the situation where there is a small set of possible stratifications, each based on a different x and the statistician is unsure which one to use. We have been assuming that the population values of an x is known for each unit in the population. But in some cases this is not really necessary. If x is a categorical variable and the population sizes of the categories are known then one can stratify the sample after it has been observed. If x is a continuous variable and certain of its population quantiles are known then these can be used to define the strata.

In post-stratification it is usually assumed that there is an auxiliary variable x which is much easier to sample than the y variable of interest. In this case

a much larger preliminary sample of the x variable is taken. This first sample is then stratified and a second smaller stratified sample is taken to observe the y values. In our approach the sample can be taken without knowing the strata membership of the units. Given the sample, our non-informative Bayesian approach allows us to weight the good stratifications higher than the poor ones where the weights do not depend on the sampling design. We have seen that the resulting estimator has good frequentist properties when the sampling design is simple random sampling as well as some other designs.

These simulations support the notion that finding a good post-stratification need not depend on the design weights. Intuitively this makes sense, since a good stratification needs to have homogeneity of the y values within strata. Stratifications with homogenous strata will tend to be better than those without this property, independently of how units are weighted within strata. Even though our method ignores the design when looking for a good stratification, at the estimation stage, the design weights could carry some additional information. One example would be if the design weights have been adjusted to account for non-response. In such cases one could do the following. First find the posterior probabilities for the stratifications ignoring the design weights. Even if the sum of these weights is the population size they will not necessarily yield the correct sum within each stratum. One could then renormalize the design weights in each stratum to sum to the stratum size and then use these new weights when calculating an estimate of the stratum total. This raises some complicated issues because the same auxiliary variables that have been used to adjust the weights to account for missing observations may have also been used to define the stratifications. This issue merits further study.

7. Final remarks

Here we have defined a non-informative (stepwise) Bayesian model which allows the statistician to consider several possible stratifications. We have shown how to select the hyperparameters that define the model in an objective fashion. Given the observed data the model yields posterior probabilities for the stratifications. This approach maintains the objectivity usually associated with design-based estimation. Guidance on computing and anecdotal evidence of our method's effective were also provided. In particular, our simulations suggest that our approach does almost as well as knowing the best stratification even when certain subgroups of the population were over sampled.

A possible criticism of our method is that the statistician may want to consider stratifications with a large (more than ten) number of possible strata. If all the stratifications have about the same number of strata this should not be a problem. But as we remarked earlier, Cochran (Section 5A.8 (Cochran, 1977)) notes that, typically, most of the precision gain from stratification is obtained with six or fewer strata. We believe that this is particularly true when a statistician is considering possible post-stratifications. This is why in our simulations we considered only a small number of stratifications with at most 8 strata.

Another issue, as we have seen in section 5, is that our model will tend to prefer stratifications with fewer strata over those with more strata. Zimmerman (2013) develops an approach that builds in a penalty term to $\Pr(h)$ that counteracts this preference. In addition he considers other sampling plans which increase the probability of getting proportional allocation across all stratifications.

Finally, we acknowledge that, in strict terms, we have defined an approach that only applies to populations where the response is categorical or very discretize (i.e. takes on a small set of possible values). However, in Section 6 we show how the statistician can obtain a posterior distribution on the set of stratifications by mapping a continuous response to set of bins, but still simulate from the posterior using the original sample of responses.

In summary, our approach can be thought of as a new kind of post-stratification. It allows a design based statistician to use an "objective" posterior distribution that lets the observed data weight a small set of pre-specified possible the stratifications. This can result in significant gains in efficiency without introducing any bias.

Acknowledgments

This research was funded in part by a doctoral dissertation fellowship from the University of Minnesota and by a dissertation fellowship from the U.S. Census Bureau.

References

Cochran, W. (1977). *Sampling Techniques (Third ed.)*. Wiley. MR0474575

Elliott, M. R. and Little, R. J. A. (2000). Model-based alternatives to trimming survey weights. *Jouirnal of Official Statistics*, 16:191–209.

Fuller, W. (2009). *Sampling Statistics*. Wiley.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman and Hall. MR1469494

Golder, P. and Yeomans, K. (1973). The use of cluster analysis for stratification. *Journal of the Royal Statistical Society, Series C*, 22:213–219.

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

Hsuan, F. (1979). A stepwise bayes procedure. *Annals of Statistics*, 7:860–868. MR0532249

Johnson, B. (1971). On admissible estimators for certain fixed sample binomial problems. *Annals of Mathematical Statistics*, 42:1579–1587. MR0418300

Little, R. J. A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88:1001–1012.

Lumley, T. (2010). *Comples Surveys, A Guide to Analysis Using R*. Wiley.

Pla, L. (1991). Determining stratum boundaries with multivariate real data. *Biometrics*, 47:1409–1422. MR1157663

R Core Team (2017). *R: A language and evironment for statistical computing*. R Foundation for Statistical Computing. http://R-project.org.

Rubin, D. B. (1981). The bayesian bootstrap. *Annals of Statistics*, 9:130–134. MR0600538

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition. MR0385023

Si, Y., Pilla, N., and Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10:605–625. MR3420817

Zimmerman, P. (2013). *Survey Sampling and Multiple Stratifications*. PhD thesis, University of Minnesota. MR3193162

Appendix A: Lemma

Here we provide a lemma needed for the Theorem presented in Section 3.

Lemma. *Let v be a fixed positive real number and k be a fixed integer greater than 1. Also, let $\mathcal{S}_{(k-1)}$ be the $(k-1)$-dimensional unit simplex, let $\mathcal{S}^0_{(k-1)}$ denote its interior, and define*

$$\mathcal{S}(\zeta)_{(k-1)} = \{p \in \mathcal{S}_{(k-1)} : p_j \geq \zeta, j = 1, 2, \ldots, k\}$$

for some $\zeta \in (0, 1/k)$. Then, assume that $q \in \mathcal{S}^0_{(k-1)}$ and that $p \in \mathcal{S}(\zeta)_{(k-1)}$. Finally, define the function

$$f_m(p, q) = \log m + \sum_{j=1}^{k} q_j \log(vp_j) + \frac{\log\Gamma(mvp_j)}{m} - \frac{\log\Gamma(m(vp_j + q_j))}{m}$$

for $m = 1, 2, \ldots$. Then, as $m \to \infty$, two results hold:

(i) f_m converges uniformly at a rate of $1/m$ to a function f on the domain $\mathcal{S}(\zeta)_{(k-1)} \times \mathcal{S}^0_{(k-1)}$ where, for a fixed p, $f(p, q)$ is a strictly convex function of q with achieves its maximum at $q = p$.

(ii) the function $d_m(p, p') = f_m(p, p) - f_m(p', p')$ converges uniformly at a rate faster than $1/m$ to zero on the domain $\mathcal{S}(\zeta)^2_{(k-1)}$.

Proof. In order to study the limiting behavior of f_m, we have to deal with the limiting behavior of the function $\log\Gamma$. Recall Stirling's formula for the Gamma function (Rudin, 1976, p.194) where $z > 0$.

$$\lim_{z\to\infty} \frac{\Gamma(z)}{\left(\frac{z-1}{e}\right)^{z-1}\sqrt{2\pi(z-1)}} = 1$$

Taking the logarithm of both sides, we can also write

$$\log\Gamma(z) = (z-1)(\log(z-1) - 1) + \frac{\log(2\pi) + \log(z-1))}{2} + a_z$$

$$= \left(z - \frac{1}{2}\right)\log(z-1) + 1 - z + \frac{\log(2\pi)}{2} + a_z$$

where $a_z \to 0$ as $z \to \infty$. We now use this result to study the limiting behavior of the function $g_m(u) = \log\Gamma(mu)/m + u - u\log(mu)$ for $u \in [v\zeta, v+1)$ as $m \to \infty$. In the lines below, we use "little-o" notation, where $o(1/m)$ refers to an error term which goes to zero more quickly than $1/m$.

$$\begin{aligned}
g_m(u) &= \log\Gamma(mu)/m + u - u\log(mu) \\
&= \frac{\left(mu - \frac{1}{2}\right)\log(mu-1) + 1 - mu + \frac{1}{2}\log(2\pi) + a_{mu}}{m} + u - u\log(mu) \\
&= \frac{-\frac{1}{2}\log(mu-1) + 1 + \frac{\log(2\pi)}{2} + a_{mu}}{m} + u\log\left(\frac{mu-1}{mu}\right) \\
&= \frac{-\log(mu-1) + \log(mu) - \log(mu) + \log(2\pi)}{2m} + \frac{a_{mu}}{m} \\
&\qquad + \frac{1 + mu\log\left(\frac{mu-1}{mu}\right)}{m} \\
&= \frac{\log(2\pi) - \log(mu)}{2m} + \frac{1 + mu\log\left(\frac{mu-1}{mu}\right)}{m} + \frac{-\log(1-\frac{1}{mu}) + 2a_{mu}}{2m} \\
&= \frac{\log(2\pi) - \log(mu)}{2m} + \frac{1 - mu\log\left(\frac{mu}{mu-1}\right)}{m} + o(1/m) \\
&= \frac{\log(2\pi) - \log(mu)}{2m} + \frac{1 - (mu-1)\log\left(\frac{mu}{mu-1}\right)}{m} \\
&\qquad + \frac{\log\left(\frac{mu}{mu-1}\right)}{m} + o(1/m) \\
&= \frac{\log(2\pi) - \log(mu)}{2m} + \frac{\log(e) - \log\left(\left(1 + \frac{1}{mu-1}\right)^{mu-1}\right)}{m} + o(1/m) \\
&= \frac{\log(2\pi) - \log(mu)}{2m} + o(1/m)
\end{aligned}$$

At this point, we can see that g_m converges to zero uniformly on $[v\zeta, v+1)$ as $m \to \infty$. Now, we apply this evaluation of g_m in studying the limit of f_m.

$$\begin{aligned}
f_m(p,q) &= \log m + \sum_{j=1}^{k} q_j \log(vp_j) + \frac{\log\Gamma(mvp_j)}{m} - \frac{\log\Gamma(m(vp_j+q_j))}{m} \\
&= \log m + \sum_{j=1}^{k} [q_j\log(vp_j) + vp_j\log(mvp_j) - vp_j + g_m(vp_j) + \\
&\qquad - (vp_j+q_j)\log(m(vp_j+q_j)) + (vp_j+q_j) - g_m(vp_j+q_j)]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{k} (vp_j + q_j)(\log(vp_j) - \log(vp_j + q_j)) + g_m(vp_j) \\
&\quad - g_m(vp_j + q_j) - q_j \\
&= -1 + \sum_{j=1}^{k} (vp_j + q_j)(\log(vp_j) - \log(vp_j + q_j)) \\
&\quad + \frac{\log\left(\frac{vp_j + q_j}{vp_j}\right)}{2m} + o(1/m)
\end{aligned}
$$

At this point, it is clear that $\lim_{m\to\infty} f_m$ exists, that f_m converges to it at a rate of $1/m$ or faster, and that it and is equal to

$$f(p,q) = -1 + \sum_{j=1}^{k} (vp_j + q_j)(\log(vp_j) - \log(vp_j + q_j))$$

We can also see that convergence at a rate of $1/m$ is uniform on the domain $(p,q) \in \mathcal{S}(\zeta)_{(k-1)} \times \mathcal{S}^0_{(k-1)}$ by looking at the error between f_m and f. Here, we use "big-O" notation, where $O(1/m)$ refers an error term that goes to zero exactly at a rate of $1/m$.

$$|f_m(p,q) - f(p,q)| < \left| \frac{k}{2m} \log\left(\frac{v+1}{v\zeta}\right) + o(1/m) \right| = O(1/m)$$

Now, we will show that f has the property described in result (i). That is, for a fixed p, $f(p,q)$ is a strictly convex function of q with its maximum at $q = p$. First, we fix $p \in \mathcal{S}(\zeta)_{(k-1)}$, and look at the partial derivative $\frac{\partial}{\partial q_j} f(p,q)$ for $j = 1, 2, \ldots, k-1$. Recall that, for $q \in \mathcal{S}^0_{(k-1)}$, q_k is actually just an abbreviation for $1 - (q_1 + q_2 + \ldots + q_{k-1})$.

$$\frac{\partial f(p,q)}{\partial q_j} = \log\left(\frac{vp_k + q_k}{vp_k}\right) - \log\left(\frac{vp_j + q_j}{vp_j}\right)$$

It is clear that, when $q = p$, all partial derivatives will equal zero. So, we only need to show that the $(k-1) \times (k-1)$-dimensional Hessian matrix is negative definite (q lies in an open set so there are no boundary conditions to consider). The second partial derivative with respect to q_j, i.e. the j^{th} diagonal element of the Hessian matrix, is

$$\frac{\partial^2 f(p,q)}{\partial q_j^2} = \frac{-1}{vp_k + q_k} + \frac{-1}{vp_j + q_j}$$

Next, the "mixed" partial derivative f with respect to some pair q_j, q'_j where $1 \le j < j' \le k$, i.e. the j, j' off-diagonal element of the Hessian matrix, is

$$\frac{\partial^2 f(p,q)}{\partial q_j q_{j'}} = \frac{-1}{vp_k + q_k}$$

which does not actually depend on j, j'. Now, if we let $\sigma = 1/(vp_k + q_k)$ and

$\tau_j = 1/(vp_j + q_j)$ for $j = 1, 2, \ldots, k-1$, the Hessian matrix is equal to $-\mathbf{A}$ where

$$\mathbf{A} = \begin{pmatrix} \sigma + \tau_1 & & \sigma \\ & \ddots & \\ \sigma & & \sigma + \tau_{k-1} \end{pmatrix}$$

and where σ and τ_j are positive for $j = 1, 2, \ldots, k-1$. Now, we only need to show that $\mathbf{A}$ is positive definite. So, let $\mathbf{D}$ be the $(k-1)$-dimensional diagonal matrix with the vector $(\tau_1, \tau_2, \ldots, \tau_{k-1})$ on the diagonal, and let $\mathbf{e}$ be the vector of 1's in $\mathbb{R}^{k-1}$. Then,

$$\begin{aligned} x^T \mathbf{A} x &= x^T \mathbf{D} x + (k-1)\sigma x^T \mathbf{Q} x \\ &\geq \sum_{j=1}^{k-1} \tau_j x_j^2 \\ &> 0 \end{aligned}$$

Therefore, $\mathbf{A}$ is positive definite, the Hessian matrix of $f(p, q)$ is negative definite, and $q = p$ minimizes $f(p, q)$ for any fixed $p \in \mathcal{S}(\zeta)_{(k-1)}$. This completes the proof of result (i). Next, we need to show that $d_m(p, p') = f_m(p, p) - f_m(p', p')$ converges uniformly to zero at a rate faster than $1/m$ on the domain $\mathcal{S}(\zeta)^2_{(k-1)}$. First, we study $f_m(p, q)$ when $p = q$.

$$\begin{aligned} f_m(p, p) &= -1 + \sum_{j=1}^{k} (v+1) p_j \log\left(\frac{v}{(v+1)}\right) + \frac{\log\left(\frac{(v+1)}{v}\right)}{2m} + o(1/m) \\ &= -1 + (v+1)\log\left(\frac{v}{(v+1)}\right) + \frac{k\log\left(\frac{(v+1)}{v}\right)}{2m} + o(1/m) \end{aligned}$$

Now, we can see that, not only does $f_m(p, p)$ converge at rate $1/m$ to a constant for $p \in \mathcal{S}(\zeta)_{(k-1)}$, but that the $O(1/m)$ term does not depend on p. Hence, $d_m(p, p') = o(1/m)$ for $p, p' \in \mathcal{S}(\zeta)_{(k-1)}$, and our proof is complete.

Appendix B: R code for computing posterior probabilities

```
#Here is an R function which computes the posterior probability
#of the possible stratifications. This assumes the uniform prior,

#maxy defines the possible values for y which are 1:maxy.
#y are values in the sample.
#H is column matrix defining  the stratifications for y values in the sample.
#lsstrsz is a list of the strata sizes.
#eps is epsilon.
#N is the popultion size. Each member of lsstrsz should sum to N.
```

```
findpostprob<-function(y,H,maxy,lsstrsz,eps,N)
    {
        K<-ncol(H)
        ans1<-rep(0,K)
        for(i in 1:K)
            {
                strsz<-lsstrsz[[i]]
                nstr<-length(strsz)
                cntmx<-NULL
                for(j in 1:nstr){
                    cntmx<-rbind(cntmx,tabulate(y[H[,i]==j],nbins=maxy))
                }
                epsvec<-(strsz/N)*eps
                for(j in 1:nstr){
                    num<-lgamma(maxy*epsvec[j]) +sum(lgamma(epsvec[j]+cntmx[j]))
                     den<-maxy*lgamma(epsvec[j]) +
                                        lgamma(epsvec[j]*maxy + sum(cntmx[j]))
                    ans1[i]<-ans1[i] + num - den
                }
            }
        fans<-rep(0,K)
         for(i in 1:K){
            fans[i]<-1/sum(exp(ans1-ans1[i]))
        }
        return(fans)
    }
```