Survey Sampling and Multiple Stratifications

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Patrick Lennon Kendall Zimmerman

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Glen Meeden, Advisor

September, 2013

Acknowledgements

- Thanks, University of Minnesota and US Census Bureau, for funding this dissertation.
- Thanks, Glen, for patiently handling my constant stream of crackpot ideas, for not letting me write anything too incomprehensible, and because your creativity sparked a lot of the main ideas in here.
- Thanks, Charlie Geyer, Lan Wang, and Tom Burk, for being willing to spend time reading and listening this stuff, and then in addition actually responding intelligently to it.
- Thanks, Maury Bramson and Joe Eaton, for helping me do some math.
- Thanks, Mom and Dad.

Dedication

To Maggie - thanks for marrying me.

Abstract

In survey sampling, stratified random sampling and post-stratification can increase the precision of estimation. In some cases, however, there may be multiple ways to stratify a population. We present a method, based on a non-informative Bayesian approach, that uses a finite mixture model to incorporate information from each stratification into estimation. This approach works well when the response variable is categorical or discrete, and for some non-response types of problems. We provide the theoretical basis for our method, present some simulation results, discuss various extensions, and define some software that implements the method.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Approaches to survey sampling

In survey sampling, our goal is to learn about some characteristic of interest, or "response variable", that takes on a value for each unit in a finite population. Let $\mathcal{U} = \{1, 2, \ldots, N\}$ index the population, and write the response variable as $y = (y_1, y_2, \ldots, y_N)$. We wish to estimate a function of y, referred to as a population parameter and denoted γ; for example, γ may be the population mean, denoted μ. This is accomplished through observing y for a sample of units, and somehow relating the observed part to the unobserved part. 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 will also denote the unsampled set of units, i.e. $\mathcal{U} - s$, with s', and the unsampled part of y with $y_{s'}$.

Survey sampling problems become statistics problems when we use a probabilistic framework as the basis for relating y_s to $y_{s'}$. A basic dichotomy splits these frameworks into design-based and model-based approaches. The design-based approach was developed in large part by researchers at the US Census Bureau (Hansen and Hurwitz, 1943),

and has been studied extensively: Cochran (1977) authored a popular textbook, and Särndal et al. (1992) provide a more recent reference. In this approach, y is regarded as fixed and unknown, and the statistical distribution of an estimator $\hat{\gamma}$ is induced by the random mechanism used to choose s, known as a sampling design (creating an indelible connection between the design and statistical inference). The pertinent theoretical properties of these estimators are of the frequentist sort, and are studied with respect to the chosen sampling design, importantly "design-consistency" where an estimator converges in probability to its target parameter as the population and sample size grow to infinity and the type of sampling design used is held constant. The fact that the statistician can often control or completely know the randomness in the problem (by controlling or knowing the sampling design) helps minimize the use of modeling assumptions, relegating their use to dealing with non-sampling errors like non-response. This emphasis on objectivity is attractive in survey sampling, which is often used to create official statistics for large-scale surveys, and it has helped the design-based approach play a dominant role in practice (Rao, 2011).

The model-based approach (which can further be split into frequentist and Bayesian approaches) has been developed more recently than the design-based approach, and is more closely related to mainstream statistics: the choice of s is typically regarded as fixed, and y is modeled as a random variable. Hence, statistical inference is not as fundamentally reliant on the sampling design. In the frequentist model-based approach (e.g. Royall, 1976, 1988; Valliant et al., 2000) y is often assumed to have some parametric distribution, indexed by a fixed and unknown θ. Then, the estimator $\hat{\gamma}$ is based on, and inherits model-based statistical properties from, a frequentist estimator of θ. In the Bayesian model-based approach (e.g. Basu, 1969; Ghosh and Meeden, 1997), a completely known prior distribution for y is chosen (i.e. if the distribution of y depends on a parameter θ, then θ must also follow some defined prior distribution), yielding a posterior distribution $y_{s'}|y_s$, and hence a distribution $\gamma|y_s$ which serves as the basis for statistical inference. Although statistical models are very effective for representing known information about a population that can dramatically improve estimation, they are also associated with subjective or hard-to-justify explicit assumptions (e.g. y is assumed to come from some parametric family of distributions, and, in the Bayesian approach, we even define a prior distribution over them). An offshoot known as the

"non-informative" Bayesian approach (e.g. Ghosh and Meeden, 1997), attempts to use formal Bayesian tools to produce a posterior distribution that combines only known, objective information about the population with information from the observed sample y_s in a posterior distribution.

The non-informative Bayesian approach is of particular interest in this thesis, but both it and the traditional subjective Bayesian approach may be unfamiliar to many readers. To provide some background, we give examples of how designed-based, Bayesian, and non-informative approaches can be used in a few common survey sampling scenarios.

1.2 Simple random sampling

In the simplest scenario, there is no information available about the population other than its size N, and we assume that a unit's index i provides no information about y_i. If we wish to estimate μ in this case, the most common course of action is the design-based simple random sampling procedure, wherein s is chosen completely at random (apart from fixing the sample size n), and we let $\hat{\mu} = \bar{y}_s$, the sample mean of y_s.

We draw heavily on the work of Ericson (1969b,a) in presenting the subjective Bayesian approach to a simple random sampling problem. Assume that y is an independent random vector where, for each unit $i \in \mathcal{U}$, y_i has the distribution

$$y_i|\theta \sim \mathrm{N}(\theta, \sigma^2)$$

where σ^2 is fixed and known. If we used a frequentist approach, we could now use frequentist methods to estimate θ based on y_s, and then predict the value of $y_{s'}$, leading to an estimator of γ. But in the Bayesian approach, we use another prior distribution to represent our uncertainty about θ,

$$\theta \sim \mathrm{N}(m, v)$$

where m and v are fixed, known hyperparameters.

It is worth pausing here to make a few notes. First, this example is rather contrived in that we assume that the variance σ^2 is known. In practice, a model like this would typically use a prior distribution to represent uncertainty about σ^2. Second, although

we have not mentioned the sampling design used along with the Bayesian model, a statistician using a model-based approach would still prefer a sampling design like simple random sampling for this scenario. Such a design will protect against concerns associated with purposive sampling and help to ensure that the sample is representative of the entire population. Finally, although contrived, this model still exemplifies a fundamental concept related to the Bayesian link between y_s and $y_{s'}$. We use a two-level hierarchical prior distribution for y, where it is a conditionally independent random vector at one level, but becomes an exchangeable random vector if we integrate across the distribution of θ. This dependence between elements of y is critical: if the unconditional prior distribution for y was as an independent and identically distributed collection of random variables, the observation of y_s would tell us absolutely nothing about $y_{s'}$. This seems intuitively wrong. Instead, when we use the exchangeable distribution presented above, y is a positively correlated collection of random variables. The distributions of y_s and $y_{s'}$ both have an a priori mean, but we are able to update our belief about the distribution of $y_{s'}$ a posteriori to make it more like what we observed in y_s. A priori exchangeability between seen and unseen units will appear throughout this thesis.

Now we show how statistical inference is actually carried out after observing y_s. We first find the posterior distribution $\theta|y_s$,

$$\theta|y_s \sim \mathrm{N}\left(\frac{(\sigma^2/n)m + v\bar{y}_s}{\sigma^2/n + v}, \frac{v\sigma^2/n}{\sigma^2/n + v}\right)$$

which can be used to determine the posterior distribution of $y_{s'}|y_s$ by noting that

$$\begin{aligned}\Pr(y_{s'}|y_s) &= \int \Pr(\theta, y_{s'}|y_s)d\theta \\ &= \int \Pr(y_{s'}|\theta)dF(\theta|y_s)\end{aligned}$$

The posterior distribution of $y_{s'}|y_s$ is then found to be

$$y_{s'}|y_s \sim \mathrm{N}\left(\frac{(\sigma^2/n)m + v\bar{y}_s}{\sigma^2/n + v}, \Sigma\right)$$

where Σ is a matrix with $\sigma^2 + \frac{v\sigma^2/n}{\sigma^2/n+v}$ on the diagonal and $\frac{v\sigma^2/n}{\sigma^2/n+v}$ on the off-diagonal. We can now use this posterior distribution, for example, to find the Bayes rule for μ

under squared-error loss:

$$\begin{aligned}\hat{\mu} &= \mathrm{E}[\mu|y_s] \\ &= \frac{1}{N}\left(\sum_{i\in s} y_i + (N-n)\mathrm{E}[y_{i'}|y_s]\right) \\ &= \bar{y}_s + \left(\frac{N-n}{N}\right)\left(\frac{\sigma^2/n}{\sigma^2/n+v}\right)(m-\bar{y}_s)\end{aligned}$$

where $i' \in s'$. Based on this same model, we can actually find estimators for an arbitrary γ, under any loss function, which is a strength of the Bayesian approach. Under squared-error loss, the estimator will simply be the posterior expectation $\mathrm{E}[\gamma|y_s]$. For example, Ericson (1969b) discusses estimation of population percentiles.

A criticism of the Bayesian approach is that the resulting estimators depend on the prior distribution used, and that this type of subjectivity is often inappropriate in survey sampling. For example, $\hat{\mu}$ is a weighted average of $\bar{y}_s$ (which is based on observed data) and m (which can be chosen at the discretion of the statistician). On the other hand, in most practical circumstances, $\hat{\mu}$ will be quite close to $\bar{y}_s$.

The non-informative Bayesian approach to a simple random sampling problems is best exemplified by the Polya posterior. The Polya urn distribution (Feller, 1968), from which the Polya posterior derives its name, goes as follows. Imagine that we have an urn containing B_1 balls labeled 1 and B_0 balls labeled 0, and a pile of R unlabeled balls sitting in a pile next to the urn. We then randomly draw a labeled ball from the urn and an unlabeled ball from the pile, apply the label from the labeled ball to the unlabeled ball, and put both into the urn. The number of balls inside the urn has now grown by one (it contains $B_1 + B_0 + 1$ balls), and the number in the pile besides the urn has shrunk by one $(R-1)$. This process is repeated until all balls from the pile have been labeled and placed in the urn. The Polya urn distribution gives the probability of observing a particular order of labels applied to the R balls placed in the urn. Let b_i be the label applied to i^{th} ball for $i = 1, 2, \ldots, R$, and let $r = (b_1 + b_2 + \ldots + b_R)$. Then,

$$\Pr(b_1, b_2, \ldots, b_R) = \frac{\Gamma(B_1+r)\Gamma(B_0+R-r)\Gamma(B_1+B_0)}{\Gamma(B_1+B_0+R)\Gamma(B_1)\Gamma(B_0)}$$

where Γ denotes the Gamma function. Although this function gives the probability an ordered set of labels, $(b_1, b_2, \ldots, b_R)$, it only depends on the sum r. Hence, any

permutation of $(b_1, b_2, \ldots, b_R)$ occurs with the same probability, making its distribution exchangeable.

The Polya posterior is a posterior distribution for $y_{s'}|y_s$ in a survey sampling problem where a unit $i \in s$ and its value y_i are analogous to a ball from the urn and its label, respectively, and a unit $i' \in s'$ whose value is unknown is analogous to an unlabeled ball from outside the urn. This analogy applies directly if there are only two distinct values observed in a population. Otherwise, the Polya urn scheme described can be generalized so that a ball from the urn can have one of any number of labels. To explicitly state this generalized posterior distribution in the survey sampling setting, let y_{zA} equal the number of units i in a set $A \subseteq \mathcal{U}$ such that $y_i = z$. Then,

$$\Pr(y_{s'}|y_s) = \frac{\Gamma(n)\prod_{z\in y_s}\Gamma(y_{z\mathcal{U}})}{\Gamma(N)\prod_{z\in y_s}\Gamma(y_{zs})}$$

Before discussing any characteristics of the Polya posterior, we should note that it does not arise from a standard Bayesian model. Rather, it results from an application of the stepwise Bayes technique, which will be discussed in Section 2.2.

Since this posterior distribution is exchangeable, just like the one from the basic Bayesian model shown above, $y_{i'}|y_s$ for each $i' \in s'$ has the same posterior distribution. In particular, the probability that $y_{i'}$ equals y_i, for any $i \in s$, is $1/n$, and its posterior expectation is $\bar{y}_s$. Hence, under squared-error loss, the estimator of μ based on the Polya posterior is

$$\begin{aligned}\hat{\mu} &= \mathrm{E}[\mu|y_s] \\ &= \frac{1}{N}\left(\sum_{i\in s} y_i + (N-n)\mathrm{E}[y_{i'}|y_s]\right) \\ &= \bar{y}_s\end{aligned}$$

One might ask why the Polya posterior is of interest given that the design-based approach already justifies the use of $\bar{y}_s$ as an estimator of μ. A theoretical benefit of the Polya posterior is that, as a stepwise Bayes rule, it implies the admissibility of $\bar{y}_s$ as an estimator of μ (see Section 2.2), a result which is not implied by the design-based approach. A practical benefit is that, like the basic Bayesian model, the Polya posterior can be used to derive an estimator of an arbitrary population parameter γ. Fortunately,

estimation of an arbitrary γ is always feasible even if a closed form for the estimator cannot be analytically derived. A Monte Carlo approximation of the posterior distribution of $y_{s'}|y_s$, obtained by simulating the Polya urn process, can be paired with the desired loss function to calculate the estimator. Also, as something of a philosophical benefit, note that, unless additional assumptions are made, the design-based approach only provides justification of $\bar{y}_s$ as an estimator of μ when simple random sampling has been used. An instance of such an "additional assumption" is seen in the case where systematic sampling has been used (e.g. s consists of each i that is a multiple of 10). In this case, the design-based properties of $\bar{y}_s$ have been justified through assumptions like "the population is in random order" (Cochran, 1977, Section 8.5). We argue that such a belief, which concerns y and not the sampling mechanism, makes strange bedfellows with the design-based approach, but is well-suited to a model-based approach. The upshot of this argument is that the Polya posterior provides a more sensible or perhaps natural basis for using $\bar{y}_s$ to estimate μ. Finally, the Polya posterior can also be favorably compared to other model-based approaches: it does not rely on a parametric assumption such as a Normal, Poisson, or other "name-brand" distribution; it also does not contain the subjectivity of a standard Bayesian approach like the model present above. Consequently, the Polya posterior carries a unique combination of attractive characteristics.

1.3 Stratification

Stratification is a common survey sampling scenario in which a population $\mathcal{U}$ can be partitioned into a disjoint collection of k subpopulations, or "strata", denoted $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_k$, with stratum means $\mu_1, \mu_2, \ldots, \mu_k$. For example, if we are estimating the mean household income in a statewide survey, we may separate the population of households into urban and rural households. When stratification is used, estimation basically consists of estimating population quantities separately for each stratum (e.g. $\hat{\mu}_j$ for $j = 1, 2, \ldots, k$), and then combining the stratum-specific estimates to form a "global" estimate (e.g. $\hat{\mu}$). This type of method is attractive because (i) the stratum-specific estimates may be of interest themselves, and (ii) the resulting global estimator will often perform better than one not based on the stratification, particularly when the within-stratum variance

of y is small.

The standard design-based method for this scenario is known as stratified random sampling (Cochran, 1977). In stratified random sampling, we choose s by selecting a random sample s_j independently for $j = 1, 2, \ldots, k$, and estimating the stratum means using the corresponding sample means. These stratum sample means, denoted $(\bar{y}_{s_1}, \bar{y}_{s_2}, \ldots, \bar{y}_{s_k})$ are then combined in a weighted average to estimate μ:

$$\hat{\mu}_{str} = \sum_{j=1}^{k} \frac{N_j}{N} \bar{y}_{s_j}$$

where N_j refers to the population size of the j^{th} stratum. This estimator can also be used from a design-based perspective even if the strata are not sampled independently, e.g. if s is chosen as a simple random sample (Holt and Smith, 1979). The resulting procedure is referred to as post-stratification.

A subjective Bayesian model can also be defined for stratified populations. If we examine the design-based stratified estimator, we can see that it keeps information about each strata separated, i.e. it reflects a belief that two units which lie in same stratum are alike, but this is not necessarily so for two units lying in different strata. A Bayesian model congruent with this belief can be constructed by assuming that y is exchangeable within strata and independent between strata. Hence, the prior distribution for y should be of the form

$$\Pr(y) = \Pr(y_{\mathcal{U}_1}) \times \Pr(y_{\mathcal{U}_2}) \times \ldots \times \Pr(y_{\mathcal{U}_k})$$

where $y_{\mathcal{U}_j}$ denotes the part of y that belongs to the j^{th} stratum. Now, we only need to specify an exchangeable prior distribution for $y_{\mathcal{U}_j}$, and can just use the basic Bayesian model from above. Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, and let h be an N-length vector where h_i gives the stratum membership of the i^{th} unit. Then,

$$y_i | (\theta_j, h_i) \overset{\text{iid}}{\sim} \text{N}(\theta_j, \sigma^2) \text{ for } i = 1, 2, \ldots, N_j; j = 1, 2, \ldots, k \tag{1.1}$$

$$\theta_j \overset{\text{iid}}{\sim} \text{N}(m, v) \text{ for } j = 1, 2, \ldots, k \tag{1.2}$$

where m and v are fixed, known hyperparameters. The Bayesian analysis then proceeds

like last time, and, conditional on the sample, we can find the posterior distribution $y_{s'}|y_s$. If we let s'_j be the unseen units in $\mathcal{U}_j$, i.e. $s'_j = \mathcal{U}_j - s_j$, then

$$y_{s'_j}|y_s \sim \mathrm{N}\left(\frac{(\sigma^2/n_j)m + v\bar{y}_{s_j}}{\sigma^2/n_j + v}, \Sigma\right) \text{ independently for } k = 1, 2, \ldots, k$$

where Σ has the same definition as above. Note that this posterior distribution is composed of an independent collection exchangeable random variables, just as in the prior distribution. We can now combine this posterior distribution with a loss function to derive Bayes rules for any parameter γ as we did above. For example, the Bayes rule for estimating μ under squared-error loss is

$$\begin{aligned}
\hat{\mu} &= \mathrm{E}[\mu|y_s] \\
&= \frac{1}{N}\sum_{j=1}^{k}\left\{n_j\bar{y}_{s_j} + (N_j - n_j)\mathrm{E}[y_i|y_{s_j}, h_i = j]\right\} \\
&= \sum_{j=1}^{k}\left\{\frac{N_j}{N}\bar{y}_{s_j} + \left(\frac{N_j - n_j}{N}\right)\left(\frac{\sigma^2/n_j}{\sigma^2/n_j + v}\right)(m - \bar{y}_{s_j})\right\}
\end{aligned}$$

We can see here again that the Bayesian estimator resembles the objective design-based estimator, but with an added bit of subjectivity that depends on the choice of the prior distribution.

Vardeman and Meeden (1984) present an extension of the Polya posterior to provide a non-informative Bayesian method of stratified estimation. The posterior distribution is basically k independent Polya posteriors, one for each stratum in the population. Specifically,

$$\Pr(y_{s'}|y_s) = \Pr(y_{s'_1}|y_{s_1}) \times \Pr(y_{s'_2}|y_{s_2}) \times \ldots \times \Pr(y_{s'_k}|y_{s_k})$$

where $y_{s'_j}$ represents the unseen units from the j^{th} stratum. Then, for each individual stratum, $j = 1, 2, \ldots, k$,

$$\Pr(y_{s'_j}|y_{s_j}) = \frac{\Gamma(n_j)\prod_{a\in y_{s_j}}\Gamma(y_{a\mathcal{U}_j})}{\Gamma(N_j)\prod_{a\in y_{s_j}}\Gamma(y_{as_j})}$$

Hence, this posterior is an independent set of k exchangeable groups of random variables, just like the posterior from the stratified Bayesian model.

The posterior also implies that, if $i \in s'_j$, for any $j \in \{1, 2, \ldots, k\}$, $\mathrm{E}[y_i|y_s] = \bar{y}_{s_j}$. The Bayes rule for estimating μ under squared-error loss is then found to be

$$\hat{\mu} = \sum_{j=1}^{k} \frac{N_j}{N} \bar{y}_{s_j}$$

which is equivalent to the stratified random sampling or post-stratification design-based estimator. Like last time, the appropriateness of the stratified Polya posterior does not depend on the sampling design, and it can be used to derive an estimator for any population parameter γ.

1.4 The Multiple Stratification approach

In the basic stratification scenario, the stratification of the population is known *a priori*. Let $h = (h_1, h_2, \ldots, h_N)$ represent a stratification, where h_i gives the stratum membership of unit i, for $i = 1, 2, \ldots, N$. This h may be defined as a one-to-one map of some completely known categorical auxiliary variable x. For instance, in the household income example described at the beginning of Section 1.3, x_i would record whether the i^{th} household was urban or rural, and h_i could equal 1 for x_i equal to "urban" and 2 for x_i equal to "rural". In other scenarios, however, the problem of defining h, or "strata formation", becomes non-trivial. The goal of strata formation is still to define a single stratification such that y has small within-stratum variance, but there may be multiple auxiliary variables available, and an auxiliary variable may be continuous instead of categorical. If there are multiple auxiliary variables, we may be able to stratify on all of them, i.e. each unique combination of their values defines a distinct stratum, but this strategy becomes unattractive and even impossible when too many auxiliary variables are known, and a subset of variables must be selected instead. For a continuous auxiliary variable x (or any variable that has too many distinct values to use as strata), some function of x must be chosen to h. For example, four strata may be defined to correspond to quartile groups of x, but other appropriate definitions could obviously be used. Because standard stratification methods (including those described in Section

1.3) assume that h is known *a priori*, strata formation can impose a challenging problem for the statistician, similar to model selection.

Various unsupervised dimension reduction techniques can be and have been used to assist in strata formation; Pla (e.g. 1991) used principal components analysis, and Golder and Yeomans (1973) used cluster analysis. However, unsupervised dimension reduction (where "unsupervised" implies that it is conducted independently of y) may not perform well when the goal of strata formation is to have small within-stratum variability in y.

Another option that avoids the problem of strata formation when there are a small number of variables but at least one continuous auxiliary variable is to based estimation on a regression model. Regression estimation can be conducted using the design-based approach (e.g. "model-assisted" estimation as in Särndal et al., 1992) or the model-based approach (e.g. Valliant et al., 2000). However, when model misspecification is of concern, forming strata may be preferable because stratification avoids the pitfalls of model misspecification while still providing the potential for efficiency gains (Fuller, 2009).[1]

In this thesis, we present a method called "multiple stratification" which allows the statistician to simultaneously consider multiple versions of h. Conceptually, multiple stratification is a very simple idea: it uses a finite mixture model where each stratification contributes a model and statistical inference averages over stratifications in some way. This could be implemented fairly simply using parametric families of distributions in a model-based approach. The desire for objectivity in survey sampling presents a more interesting problem, however, and we use a non-informative Bayesian approach in order to combine the finite mixture model concept with the objectivity of design-based estimators.

An outline of the rest of this dissertation is as follows. In Chapter 2, we present a multiple stratification model and estimator. In Chapter 3, we extend the multiple stratification method to handle problems with non-response. In Chapter 4, we provide functions that can implement our methods in the R statistical computing language (R Core Team, 2013).

[1] For an interesting combination of regression and stratification techniques, consider the endogenous post-stratification method described in Breidt (2008) and Dahlke et al. (2013).

Chapter 2

A Multiple Stratification Estimator

In this chapter, we present a method for multiple stratification estimation using a non-informative Bayesian approach. In Section 2.1, we present this method for the special case when y is a vector of binary variables. This allows us to explain most of our approach without having to use the stepwise Bayes technique. In Section 2.2, we extend the method to apply for an arbitrary parameter space for y. We then explain our choice for some hyperparameters that appear in our models (Section 2.3), describe how stratifications with differing numbers of strata can be handled (Section 2.4), present a sampling mechanism that works well when paired with multiple stratification estimation (Section 2.5), and a discuss a Bayesian version of sampling weights based on multiple stratifications (Section 2.6). Finally, in Section 2.7, some simulated examples are presented.

2.1 Estimation for a binary response

Here, we handle the case where $y \in \{0,1\}^N$. Although we use a proper prior distribution, we still take a non-informative approach: our prior distribution $\Pr(y)$ is not meant to represent subjective prior beliefs about the population. 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

$i = 1, 2, \ldots, N$. Also, let k_h denote the number of strata defined by $h \in \mathcal{H}$. We assume that a prior distribution over $\mathcal{H}$ is known. Although, in some cases, a statistician may be comfortable defining $\Pr(h)$ across $\mathcal{H}$ based on subjective belief or past performance of the stratifications (the latter being more in tune with a non-informative approach), we acknowledge that this will not always be true. As a default, it should typically be appropriate to set $\Pr(h)$ to be a constant for all $h \in \mathcal{H}$ such that $k_h = k$, for each possible k. Section 2.4 explains how to deal with a varying k_h. Finally, we will attach, as a subscript, the index pair jh to a set A when referring to the part of A that lies in stratum j according to h for $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$. This defines 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.

Now, we can define the prior distribution for y. First, we will lay out the general structure of the distribution.

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

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 $y_{\mathcal{U}_{1h}}, y_{\mathcal{U}_{2h}}, \ldots, y_{\mathcal{U}_{k_h h}}$.

To complete the definition of $\Pr(y)$, we need to define $\Pr(y_{\mathcal{U}_{jh}}|h)$ for each $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$. We do this by using a Beta-Binomial type of model with a stratum-specific "process mean" variable.

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

where ϵ_{jh} for $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$ are known hyperparameters. In keeping with a non-informative approach, these hyperparameters should be chosen to reflect

known, objective information about the population. Specifically, we recommend setting $\epsilon_{jh} = \epsilon N_{jh}/N$ where ϵ is a small number, e.g. 1/10 or 1/100. This recommendation is explained in Section 2.3.

We can integrate across θ_{jh} to obtain a concise expression of $\Pr(y_{\mathcal{U}_{jh}}|h)$. Let y_{zA} denote that number of units i in a set $A \subseteq \mathcal{U}$ where $y_i = z$. Then,

$$\begin{aligned}\Pr(y_{\mathcal{U}_{jh}}|h) &= \frac{\Gamma(2\epsilon_{jh})}{\Gamma(\epsilon_{jh})^2}\int_0^1 \theta^{\epsilon_{jh}+y_{1\mathcal{U}_{jh}}-1}(1-\theta)^{\epsilon_{jh}+y_{0\mathcal{U}_{jh}}-1}d\theta \\ &= \frac{\Gamma(2\epsilon_{jh})\Gamma(\epsilon_{jh}+y_{1\mathcal{U}_{jh}})\Gamma(\epsilon_{jh}+y_{0\mathcal{U}_{jh}})}{\Gamma(\epsilon_{jh})^2\Gamma(2\epsilon_{jh}+N_{jh})}\end{aligned}$$

Note that $\Pr(y_{\mathcal{U}_{jh}}|h)$ does not depend on the order of units in $y_{\mathcal{U}_{jh}}$, implying that it is an exchangeable distribution.

A convenient characteristic of $\Pr(y)$ is that the marginal probability function for a subset of y, e.g. $\Pr(y_s)$, has a similar form. First, the structure associated with the finite mixture model and the conditional independence between strata is preserved.

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

Second, the form of $\Pr(y_{s_{jh}}|h)$ is analogous to that of $\Pr(y_{\mathcal{U}_{jh}}|h)$.

$$\begin{aligned}
\Pr(y_{s_{jh}}|h) &= \sum_{y'_{\mathcal{U}_{jh}}:y'_{s_{jh}}=y_{s_{jh}}} \Pr(y'_{\mathcal{U}_{jh}}|h) \\
&= \sum_{y'_{\mathcal{U}_{jh}}:y'_{s_{jh}}=y_{s_{jh}}} \frac{\Gamma(2\epsilon_{jh})}{\Gamma(\epsilon_{jh})^2} \int_0^1 \theta^{\epsilon_{jh}+y'_{1\mathcal{U}_{jh}}-1}(1-\theta)^{\epsilon_{jh}+y'_{0\mathcal{U}_{jh}}-1} d\theta \\
&= \frac{\Gamma(2\epsilon_{jh})}{\Gamma(\epsilon_{jh})^2} \int_0^1 \sum_{y'_{\mathcal{U}_{jh}}:y'_{s_{jh}}=y_{s_{jh}}} \theta^{\epsilon_{jh}+y'_{1\mathcal{U}_{jh}}-1}(1-\theta)^{\epsilon_{jh}+y'_{0\mathcal{U}_{jh}}-1} d\theta \\
&= \frac{\Gamma(2\epsilon_{jh})}{\Gamma(\epsilon_{jh})^2} \int_0^1 \theta^{\epsilon_{jh}+y_{1s_{jh}}-1}(1-\theta)^{\epsilon_{jh}+y_{0s_{jh}}-1} \beta_\theta(y_{s_{jh}}) d\theta
\end{aligned}$$

where

$$\beta_\theta(y_{s_{jh}}) = \sum_{y'_{\mathcal{U}_{jh}}:y'_{s_{jh}}=y_{s_{jh}}} \theta^{y'_{1s'_{jh}}}(1-\theta)^{y'_{0s'_{jh}}}$$

Now, we can use the Binomial Theorem to simplify this expression

$$\begin{aligned}
\beta_\theta(y_{s_{jh}}) &= \sum_{y'_{\mathcal{U}_{jh}}:y'_{s_{jh}}=y_{s_{jh}}} \theta^{y'_{1s'_{jh}}}(1-\theta)^{y'_{0s'_{jh}}} \\
&= \sum_{c=0}^{N_{jh}-n_{jh}} \binom{N_{jh}-n_{jh}}{c} \theta^c (1-\theta)^{N_{jh}-n_{jh}-c} \\
&= (\theta + (1-\theta))^{N_{jh}-n_{jh}} \\
&= 1
\end{aligned}$$

Then,

$$\begin{aligned}
\Pr(y_{s_{jh}}|h) &= \frac{\Gamma(2\epsilon_{jh})}{\Gamma(\epsilon_{jh})^2} \int_0^1 \theta^{\epsilon_{jh}+y_{1s_{jh}}-1}(1-\theta)^{\epsilon_{jh}+y_{0s_{jh}}-1} d\theta \\
&= \frac{\Gamma(2\epsilon_{jh})\Gamma(\epsilon_{jh}+y_{1s_{jh}})\Gamma(\epsilon_{jh}+y_{0s_{jh}})}{\Gamma(\epsilon_{jh})^2\Gamma(2\epsilon_{jh}+n_{jh})}
\end{aligned}$$

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,

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

This implies that the posterior distribution also maintains the general structure of the prior distribution. First, it is a finite mixture model where each possible stratification $h \in \mathcal{H}$ supplies a different model for the unseen units, i.e. $y_{s'}|y_s, h$. Second, the unseen units from different strata are conditionally independent given some $h \in \mathcal{H}$. Next, we examine the conditional posterior distribution within a stratum.

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

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 $z \in \{0,1\}$,

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

As $\epsilon_{jh} \to 0$, this marginal distribution approaches the empirical distribution implied by observing $y_{s_{jh}}$, i.e. $\Pr(y_i = z|y_s,h) = y_{zs_{jh}}/n_{jh}$. To take this example one step farther, we can calculate the conditional posterior expectation of the same unseen unit,

$$\begin{aligned}\mathrm{E}[y_i|y_s,h] &= \sum_{z=0}^{1} z\,\Pr(y_i = z|(y_s,h)) \\ &= \frac{n_{jh}}{2\epsilon_{jh}+n_{jh}}\bar{y}_{s_{jh}} + \frac{\epsilon_{jh}}{2\epsilon_{jh}+n_{jh}}\end{aligned}$$

where $\bar{y}_{s_{jh}}$ is the mean of $y_{s_{jh}}$. We can see here that, as $\epsilon_{jh} \to 0$, $\mathrm{E}[y_i|y_s,h] \to \bar{y}_{s_{jh}}$.

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 $\Pr(y_s|h)$ is how the observed data help determine the mixture weights. In particular, $\Pr(y_s|h)$ 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 choice of hyperparameters, and we will discuss this in Section 2.3.

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 example, when $\gamma(y)$ is a linear function of y, the estimator is

$$\mathrm{E}[\gamma(y)|y_s] = \sum_{h\in\mathcal{H}} \Pr(h|y_s)\gamma\left(\mathrm{E}[y|y_s,h]\right)$$

More specifically, if $\gamma(y)$ is taken to be the population mean, denoted μ, the Bayes rule is

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

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

2.2 Estimation for a general response

The stepwise Bayes technique

Before using the stepwise Bayes technique, we provide some background. Johnson (1971) presented a special case of the 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. The technique has been used in survey sampling problems to provide a theoretical justification for frequentist estimators without appealing to the design-based approach, and to provide guidance in situations where a design-based approach is awkward to apply (e.g. Ghosh and Meeden, 1997). For instance, a stepwise Bayes justification of the design-based stratified estimator has been provided by Vardeman and Meeden (1984). In fact, the stepwise Bayes model that we present below is related to some of Vardeman and Meeden's work. But first we explain the basic stepwise Bayes concept in a general decision-theoretic setting, more or less paraphrasing Hsuan (1979). A characteristic of the stepwise Bayes technique is the use of an ordered collection of an arbitrary number of prior distributions. In the interest of brevity, though, we explain a case with just two prior distributions. The basic concept remains the same.

Suppose we are to estimate the parameter θ based on an observation of some data z whose distribution is parameterized by θ, i.e. $z \sim p_\theta(z)$ where p_θ can be a probability density or mass function. Let Θ be the parameter space and π_1 be a prior distribution on it, so we can write the Bayes risk of a decision rule δ as $r(\pi_1, \delta)$. Now, if the support π_1 is some proper subset of Θ, say Θ_1, it is possible that π_1 will not imply a unique Bayes rule. Suppose that $D(\pi_1)$ is the set of decision rules that minimize the Bayes risk under π_1. Then the idea of the stepwise technique is to obtain a unique rule from $D(\pi_1)$ by using a second prior, say π_2. The stepwise Bayes rule is then defined as

$$\delta_{SB} = \text{arginf}_{\delta \in D(\pi_1)} r(\pi_2, \delta) \tag{2.1}$$

Note that δ_{SB} is not necessarily Bayes with respect to π_2 when that prior is considered on its own. To provide one more concrete detail of this technique, all rules within $D(\pi_1)$ will be equal on the set

$$\mathcal{Z}_1 = \{z : p_\theta(z) > 0 \text{ for some } \theta \in \Theta_1\}$$

Hence, the image of $\delta_{SB}(z)$ is determined by π_1 for $z \in \mathcal{Z}_1$ and by π_2 for $z \notin \mathcal{Z}_1$.[1]

A fundamental characteristic of a unique stepwise Bayes rule is admissibility. First, we note that, for the stepwise Bayes rule described here to be unique, it must uniquely satisfy Equation 2.1. Now, to understand why such a rule is admissible, we provide a sketch of the proof given by Hsuan (1979), which proof is really just an extension of the proof that a unique Bayes rule is admissible. Suppose that δ_{SB} is a unique stepwise Bayes rule on the ordered prior distributions (π_1, π_2) but is dominated by δ', i.e. δ' is at least as good as δ_{SB} (i.e. not have a larger risk function) on all of Θ and be better (i.e. have a smaller risk function) for some $\theta \in \Theta$. Since δ_{SB} is Bayes with respect to π_1, δ' must be equivalent to it on $\mathcal{Z}_1$ as defined above. But then Equation 2.1 also implies that the rules must be equivalent on $\mathcal{Z} - \mathcal{Z}_1$, and hence $\delta' = \delta_{SB}$, contradicting our supposition.

[1] If using π_2 to choose between rules in $D(\pi_1,)$ still does not define a unique Bayes rule for all $z \in \mathcal{Z}$, we can iterate the procedure of defining another prior distribution. This can be done an arbitrary number of times.

A stepwise Bayes multiple stratification model

Now, we use the stepwise Bayes technique to handle the case where the response is not binary. Assume that $B = \{b_1, b_2, \ldots, b_A\}$, for some positive integer A, is the set of values which can be taken by y_i for any $i \in \mathcal{U}$. We will let $B = \{1, 2, \ldots, A\}$ for convenience, but an element $b \in B$ could be a real number, a category, or anything else. The only important thing is that B is a finite set. Later, we will show that our estimation method still produces admissible estimators even if B is infinite, but for now the parameter set in which y lies is $\mathcal{Y} = B^N$.

As described above, we will use an ordered collection of prior distributions to define estimation for disjoint subsets of $\mathcal{Y}$, which we refer to as "restricted parameter sets". The ordering of the distributions can be organized into A phases, with $\binom{A}{a}$ prior distributions occurring in the a^{th} phase for $a = 1, 2, \ldots, A$. Although the phases must precede in a given order, the distributions within each phase does not matter and can proceed in an arbitrary order.

In the first phase, there is one prior distribution and corresponding restricted parameter set for each element in B. To help define them, let unique(y) be the set of distinct values from B which appear in a given $y \in \mathcal{Y}$. Then, for each $b \in B$, we define the restricted parameter set

$$\mathcal{Y}_{1b} = \{y \in \mathcal{Y} : \text{unique}(y) = b\}$$

That is, every unit in the population has a response equal to b. Hence, the restricted parameter sets in the first phase are

$$\mathcal{Y}_{11}, \mathcal{Y}_{12}, \ldots, \mathcal{Y}_{1A}$$

and each has the corresponding trivial prior distribution which puts all probability mass on its single element. A posterior distribution associated with this phase simply implies that, if $y_i = b$ for every $i \in s$, then $y_i = b$ for every $i \in s'$, too, making estimation trivial.

Next, we can handle the second through the A^{th} phrase at once by defining the prior distributions associated with the a^{th} phase, for $a \in \{2, 3, \ldots, A\}$. Each of the $\binom{A}{a}$ prior distributions in this phase correspond to a unique combination of a elements from B. For such a unique combination, b, we have the restricted parameter set

$$\mathcal{Y}_{ab} = \{y \in \mathcal{Y} : \text{unique}(y) = b\}$$

Note that the restricted parameter sets from the first phase also follow this definition. Based on the fact that a restricted parameter set is the support of a prior distribution, we can now see that every $y \in \mathcal{Y}$ receives positive probability from exactly one prior distribution: for every $y \in \mathcal{Y}$, there is exactly one $\mathcal{Y}_{ab}$ where unique$(y) = b$, and $\mathcal{Y}_{ab}$ is the support of exactly one prior distribution in the collection.

Now, for notational convenience, let $b = \{1, 2, \ldots, a\}$. Then, the prior distribution on $\mathcal{Y}_{ab}$ has the same general structure as seen in the model we used for a binary response.

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

The conditional distribution $\Pr(y_{\mathcal{U}_{jh}}|h)$ is also analogous to the binary response version. For each $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$,

$$\Pr(y_{\mathcal{U}_{jh}}|h) = \frac{\Gamma(a\epsilon_{jh}) \prod_{z=1}^{a} \Gamma(\epsilon_{jh} + y_{z\mathcal{U}_{jh}})}{\Gamma(\epsilon_{jh})^a \Gamma(a\epsilon_{jh} + N_{jh})}$$

As in the binary response section, $\Pr(y_{\mathcal{U}_{jh}}|h)$, is an exchangeable distribution because it does not depend on the ordering of y within $\mathcal{U}_{jh}$, and it is a Dirichlet-Multinomial (Beta-Binomial when $a = 2$) type of distribution where the Dirichlet mixing parameter has been integrated out. As we showed in Section 2.1, $\Pr(y_{\mathcal{U}_{jh}}|h)$ has the property that the distribution for a subset of units will have the same form. In particular, the conditional prior distribution of the sample, $y_s|h$, is

$$\Pr(y_s|h) = \prod_{j=1}^{k_h} \frac{\Gamma(a\epsilon_{jh}) \prod_{z=1}^{a} \Gamma(\epsilon_{jh} + y_{zs_{jh}})}{\Gamma(\epsilon_{jh})^a \Gamma(a\epsilon_{jh} + n_{jh})}$$

At this point, we can see concretely how the stepwise Bayes analysis will proceed under this model. Recall that, after y_s is observed, the first prior distribution in the ordered collection which assigns positive probability to y_s is selected, and, from that point, a standard Bayesian analysis is carried out. For this model, that first prior distribution will always be the one defined over $\mathcal{Y}_{ab}$ where $b = \text{unique}(y_s)$. Hence, use of this model will include the assumption that the values appearing in $y_{s'}$ are the same as those observed in y_s.

The general structure for the posterior distribution, $y_{s'}|y_s$, is

$$\Pr(y_{s'}|y_s) = \sum_{h\in\mathcal{H}} \Pr(h|y_s) \prod_{j=1}^{k_h} \Pr(y_{s'_{jh}}|y_{s_{jh}}h)$$

where, given $h \in \mathcal{H}$, the conditional posterior probability for the unseen units in the j^{th} stratum is

$$\Pr(y_{s'_{jh}}|y_{s_{jh}}, h) = \frac{\Gamma(a\epsilon_{jh} + n_{jh}) \prod_{z=1}^{a} \Gamma(\epsilon_{jh} + y_z u_{jh})}{\Gamma(a\epsilon_{jh} + N_{jh}) \prod_{z=1}^{a} \Gamma(\epsilon_{jh} + y_{z s_{jh}})}$$

Then, given $h \in \mathcal{H}$, the conditional expected value of an unseen unit $i \in s'_{jh}$ is

$$\begin{aligned} \mathrm{E}[y_i|y_s, h] &= \sum_{z=1}^{a} z \Pr(y_i = z|(y_s, h)) \\ &= \frac{n}{a\epsilon_{jh} + n_{jh}} \bar{y}_{s_{jh}} + \frac{\epsilon_{jh}}{a\epsilon_{jh} + n_{jh}} \sum_{z=1}^{a} z \end{aligned}$$

As $\epsilon_{jh} \to 0$, $\mathrm{E}[y_i|y_s, h]$ will approach $\bar{y}_{s_{jh}}$.

Finally, we discuss the admissibility of estimators produced by this stepwise Bayes model. Recall that we assumed the existence of a finite set B containing all possible values that a unit y_i could take, for $i \in \{1, 2, \ldots, N\}$, but placed no restrictions on the type of elements B could contain. We also showed that, under this set-up, every $y \in \mathcal{Y}$ receives positive probability from one prior distribution, yielding unique Bayes rules under squared-error loss. So, for any multiple stratification survey sampling problem where y can take on only a finite number of distinct values, our model produces admissible estimators. However, we are also able to produce admissible estimators when y can take on an infinite number of distinct values. For example, suppose that $y \in \mathbb{R}^N$. Because of the lack of restrictions on B, we can say that an estimator, δ, based on our stepwise Bayes model is admissible when $\mathbb{R}^N$ is reduced to any finite subset containing y. This property is known as finite admissibility, and it actually implies that the same estimator is admissible for the full parameter space of $\mathbb{R}^N$. To see why, assume that another estimator, δ', dominates it. Then δ' must be better than δ on at least some part of $\mathbb{R}^N$ and at least good on the rest of it. But this cannot be true because we can

show that δ is admissible on any finite number of parameter points where δ' supposedly dominates δ. Therefore, our estimation method can be used to produce admissible estimators for any parameter space, finite or infinite.

2.3 Choosing hyperparameters

In the models presented above, a set of hyperparameters $\epsilon_{1h}, \epsilon_{2h}, \ldots, \epsilon_{k_h h}$ is associated with each stratification $h \in \mathcal{H}$. In total, this implies that $\sum_{h \in \mathcal{H}} k_h$ hyperparameters must be defined, which may seem overwhelming. However, there is non-informative way to choose the hyperparameters based on the when we belief, given any $h \in \mathcal{H}$, that (i) inference based on $y|y_s, h$ should agree with the standard design-based estimators for stratified random sampling or post-stratification (or the estimator given in Vardeman and Meeden (1984)), and (ii) proportional allocation is a good sample allocation.[2] For now, we will assume that k_h equals some constant k for every $h \in \mathcal{H}$, and we show how to adjust for a varying k_h later on. Given this assumption and our two beliefs, we recommend setting $\epsilon_{jh} = \epsilon N_{jh}/N$ for each $h \in H$ and $j = 1, 2, \ldots, k_h$ where ϵ is on the order of $1/10$ or $1/100$.

Our recommended choice of hyperparameters is based on separately considering how the choice affects $y|(y_s, h)$ and $h|y_s$. As the reader may have picked up above, for any $h \in \mathcal{H}$, choosing ϵ_{jh} to be small for $j = 1, 2, \ldots, k$ will make inference conditional on h agree with design-based (post)-stratification. The relative sizes of different hyperparameters is not important in this regard, so long as they are all small. So, ϵ_{jh} should be "small", but, of course, exactly how small they need to be is somewhat a matter of personal preference. In the simulated examples we have considered (Section 2.7), using $\epsilon = 1/10$ seems sufficient, i.e. setting the hyperparameters any smaller produces negligible changes to inference.

Now, we consider relationship between hyperparameter choice and the distribution $h|y_s$. 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

[2] The guidance offered in this section could easily be used when some other Neyman-optimal allocation (Cochran, 1977) can be defined. We work with proportional allocation because it represents a default "naive" Neyman-optimal allocation for use when the statistician is unwilling to make assumptions about the relative size of stratum-specific variances

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 hyperparameter choice. 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, in a rough sense to be specified later, 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.

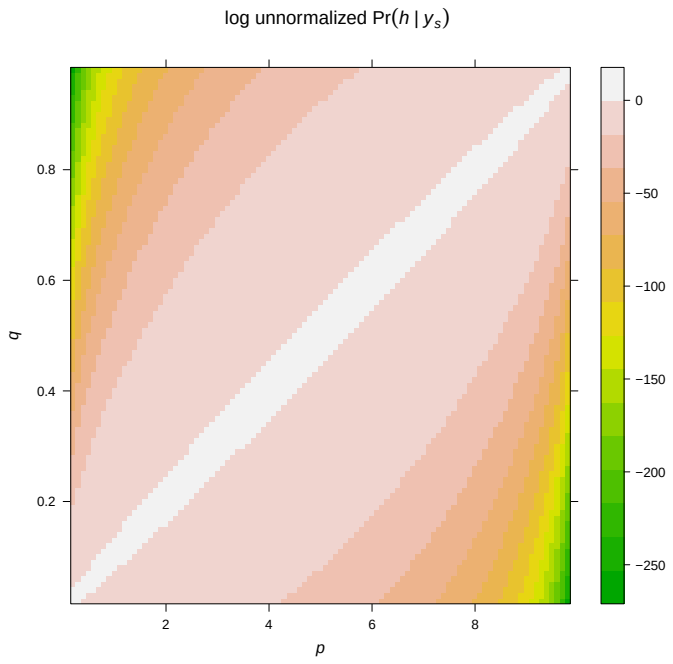

Figure 2.1: Evaluated surface of $\log \Pr(h|y_s)$ the population and sample stratum sizes defined by some h where $k_h = 2$. On the axes, $p = N_{1h}/N$ and $q = n_{1h}/n$.

In order to study $\Pr(h|y_s)$ as it relates to sample allocation, we consider a scenario where $\Pr(h)$ is uniform and y_s is fixed so that y_i is a distinct value for each $i \in s$. When $\Pr(h)$ is not uniform or not all the units in y_s have distinct values, hyperparameter

choice will still affect the distribution $h|y_s$ by creating a preference for certain sample allocations, and our recommended choice will still create a sensible preference. However, focusing on this scenario makes it easier to see how the relationship between allocation and $h|y_s$ works and to recommend a choice.

First, we present an example to help orient the reader for a theoretical analysis. Suppose that $k = 2$, that $n = 100$, and that $\epsilon_{jh} = N_{jh}/N$ for some $h \in H$. 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 $q = n_{1h}/n$. In Figure 2.1, we present an evaluated grid surface of an unnormalized $\log \Pr(h|y_s)$ over every pair $(p, q) \in \{0.02, 0.03, \ldots, 0.097, 0.098\}^2$. The simplest description of this surface is 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 seem 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 $p = q$.

We now show that these two evident properties are approximately true (for any k) when n is large. First, in a lemma, we study some properties of a function related to $\Pr(h|y_s)$.

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

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

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

$$\begin{aligned}\log\Gamma(z) &= (z-1)(\log(z-1)-1) + \frac{\log(2\pi)+\log(z-1))}{2} + a_z \\ &= \left(z-\frac{1}{2}\right)\log(z-1) + 1 - z + \frac{\log(2\pi)}{2} + a_z\end{aligned}$$

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 $u > 0$ 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 $1/m$.

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

At this point, we can see that $g_m(u)$ converges to zero uniformly on any fixed strictly positive interval. Now, we apply this evaluation of g_m in studying the limit of f_m.

$$
\begin{aligned}
f_m(p,q) &= \log m + \sum_{j=1}^{k} q_j \log(vp_j) + \frac{\log\Gamma(mvp_j)}{m} - \frac{\log\Gamma(m(vp_j+q_j))}{m} \\
&= \log m + \sum_{j=1}^{k} \left[q_j\log(vp_j) + vp_j\log(mvp_j) - vp_j + g_m(vp_j) + \right. \\
&\qquad \left. -(vp_j+q_j)\log(m(vp_j+q_j)) + (vp_j+q_j) - g_m(vp_j+q_j)\right] \\
&= \sum_{j=1}^{k} (vp_j+q_j)(\log(vp_j) - \log(vp_j+q_j)) + g_m(vp_j) - g_m(vp_j+q_j) - q_j \\
&= -1 + \sum_{j=1}^{k} (vp_j+q_j)(\log(vp_j) - \log(vp_j+q_j)) + \frac{\log\left(\frac{vp_j+q_j}{vp_j}\right)}{2m} + o(1/m)
\end{aligned}
$$

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 $1 \leq j < j' \leq k$, 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,

$$\begin{aligned}\mathbf{A} &= \mathbf{D} + \sigma \mathbf{e}\mathbf{e}^T \\ &= \mathbf{D} + (k-1)\sigma \mathbf{Q}\end{aligned}$$

where $\mathbf{Q} = \mathbf{e}\mathbf{e}^T/(k-1)$. It is easy to see that $\mathbf{Q}$ is a projection matrix by checking that it is symmetric and idempotent. Hence, it is non-negative definite. Now, let x be any non-zero vector in $\mathbb{R}^{k-1}$, and 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 $p \in \mathcal{S}'_{(k-1)}$. 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.

Now, before stating our result based on this lemma, we will take a moment to explain how we do asymptotics 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 $y = \{y_i\}_{i=1}^{\infty}$ for $N = 1, 2, \ldots$. Also, let $\mathcal{H}$ be a finite set of infinite sequences that stratify y into k strata, and let H_N be the set that contains, for each sequence $h \in \mathcal{H}$, a vector h_N of the first N terms from h. 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 $h \in \mathcal{H}$ may not appear in h_N for small values of N, but we ignore this problem because, with only finitely stratifications in $\mathcal{H}$, 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 $p_{Njh} = N_{jh}/N$, 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, ϵ_{Njh}, be set equal to ϵp_{Njh}, 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 $q_{Njh} = n_{Njh}/n_N$ where n_{Njh} 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, ignore this problem, too, because it will not be issue once N reaches some finite threshold. Finally, write $p_{Nh} = (p_{N1h}, p_{N2h}, \ldots, p_{Nkh})$ and $q_{Nh} = (q_{N1h}, q_{N2h}, \ldots, q_{Nkh})$.

Theorem. Assume that, for each $h \in \mathcal{H}$, there is some fixed $p_h, q_h \in \mathcal{S}^0_{(k-1)}$ such that $p_{Nh} \to p_h$ and $q_{Nh} \to q_h$ as $N \to \infty$. Also, assume that y 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 $h, h' \in \mathcal{H}$ 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 $h, h' \in \mathcal{H}$, $p_h = q_h$ 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 $N = 1, 2, \ldots$,

$$\begin{aligned}
\frac{\log \Pr(h_N|y_{s_N})}{n_N} &\propto \frac{1}{n_N} \sum_{j=1}^{k} \log \left(\frac{\Gamma(n_N \epsilon p_{Njh}) \Gamma(\epsilon p_{Njh} + 1)^{n_{Njh}} \Gamma(\epsilon p_{Njh})^{(n_N - n_{Njh})}}{\Gamma(\epsilon p_{Njh})^{n_N} \Gamma(n_N \epsilon p_{Njh} + n_{Njh})} \right) \\
&\propto \frac{1}{n_N} \sum_{j=1}^{k} \log \left(\frac{\Gamma(n_N \epsilon p_{Njh}) (\epsilon p_{Njh})^{n_{Njh}}}{\Gamma(n_N \epsilon p_{Njh} + n_{Njh})} \right) \\
&\propto \sum_{j=1}^{k} q_{Njh} \log(\epsilon p_{Njh}) + \frac{\log \Gamma(n_N \epsilon p_{Njh})}{n_N} - \frac{\log \Gamma(n_N(\epsilon p_{Njh} + q_{Njh}))}{n_N} \\
&\propto f_{n_N}(p_{Nh}, q_{Nh})
\end{aligned}$$

where f_m is the function from the lemma if we choose the ϵ for the asymptotic set-up and the v from the lemma to be the same. So, for any pair $h, h' \in \mathcal{H}$ and any $N \in \{1, 2, \ldots\}$,

$$\log \Pr(h_N|y_{s_N}) - \log \Pr(h'_N|y_{s_N}) = n_N (f_{n_N}(p_{Nh}, q_{Nh}) - f_{n_N}(p_{Nh'}, q_{Nh'}))$$

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

$$|f_{n_N}(p_{Nh}, q_{Nh}) - f_{n_N}(p_{Nh'}, q_{Nh'}) - (f(p_h, q_h) - f(p_{h'}, q_{h'}))| < \delta$$

Now, if h and h' 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 $f_{n_N}(p_{Nh}, q_{Nh}) < f_{n_N}(p_{Nh'}, q_{Nh'})$ 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 $h, h' \in \mathcal{H}$ that fit the scenario described for result (ii),

$$\log \Pr(h_N|y_{s_N}) - \log \Pr(h'_N|y_{s_N}) = n_N d_{n_N}(p_{Nh}, p_{Nh'})$$

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 $\{(p_h, p_{h'} : h, h' \in \mathcal{H}\}$. Hence, we can find N_2 such that, for any $h, h' \in \mathcal{H}$ 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 completes our proof.

■

This theorem has essentially shown that the properties evident in Figure 2.1 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 $\Pr(h|y_s)$.

2.4 Accounting for differing numbers of strata

Dealing with differing numbers of strata is a hard problem. Consider the relationship between this problem and selecting the correct number of clusters in a cluster analysis (e.g. Fraley and Raftery, 1998). It is an important problem, though: our multiple stratification model would not be very useful if the posterior probability of a poor-fitting stratification dominates that of a good-fitting stratification just because poor-fitting stratification defines more or fewer strata than the other. We do not purport to have completely solved this problem here, but we can offer a strategy that seems reasonable and works well in at least one example (Section 2.7). This strategy will be thoroughly *non-informative* Bayesian: it involves choosing a prior distribution based on a frequentist model-based approach, not on subjective beliefs. We also should note that it was partly inspired by the "gap statistic" proposed by Tibshirani et al. (2001) for determining the number of clusters in a data set.

Here is the reasoning behind our strategy. Suppose that we have the following set of conditions, which we will call "baseline" conditions: we have no *a priori* preference for any particular stratification $h \in \mathcal{H}$, y is unfortunately not related to any $h \in \mathcal{H}$, and simple random sampling is used to select s so that no particular stratification is likely to get a better sample allocation than another. Under these baseline conditions, we claim that, although sampling error may produce some non-uniformity in $\Pr(h|y_s)$ for particular samples, it is irrational for the expected value of the posterior distribution $\Pr(h|y_s)$ upon repeated sampling of s to be non-uniform on $\mathcal{H}$. In other words, under these baseline conditions, we believe there is no reason to prefer a particular h, so any non-uniformity in $\Pr(h|y_s)$ upon repeated sampling is unwanted. Hence, we propose that a sensible approach to accounting for a varying number of strata is to choose $\Pr(h)$ so that, when these baseline conditions hold, the expected value of the posterior distribution $\Pr(h|y_s)$ is uniform on $\mathcal{H}$.[3]

To be more explicit, suppose that we must define a prior distribution over $\mathcal{H}$ where $K_{\mathcal{H}}$ is the set of k_h for all $h \in \mathcal{H}$, and define M to be a model that instantiates our baseline conditions. In the model M, we define a new set of stratifications $\mathcal{H}_M$ such that $\mathcal{H}_M$ contains exactly one stratification with k strata for each $k \in K_{\mathcal{H}}$, and that the strata defined by every $h \in \mathcal{H}_M$ are approximately equally sized (approximately because of the finite size of $\mathcal{U}$). Also in the model M, let Y be a random vector of length N whose distribution does not depend on $\mathcal{H}_M$, and let $\tilde{s}$ be a simple random sample of $\mathcal{U}$. We'll say more about the exact distribution of Y below. Now, for $h \in \mathcal{H}_M$,

$$\mathrm{E}_M[\Pr(h|Y_{\tilde{s}})] = \Pr(h)\mathrm{E}_M[\Pr(Y_{\tilde{s}}|h)]$$

where $\Pr(Y_{\tilde{s}}|h)$ is calculated using the hyperparameter choices recommended in Section 2.3 as applied to $\mathcal{H}_M$. So, if $\mathcal{H}_M$ were the set of stratifications for our actual problem, our line of reasoning implies that, for $h \in \mathcal{H}_M$, we must define

$$\Pr(h) \propto \mathrm{E}_M[\Pr(Y_{\tilde{s}}|h)]^{-1}$$

[3] This idea has something in common with the choosing prior distributions that are invariant under some types of reparameterizations (e.g. Jeffreys, 1946). In that problem, the statistician reasons that, if inference should not depend on such a reparameterization, then she should choose a prior distribution so that it does not.

Although this expectation may be hard to calculate in closed form, it is easy to compute a Monte Carlo approximation in practice.

Now, what does this tell us about defining a prior distribution over $\mathcal{H}$, which is what we actually care about? After all, the model M is rather contrived and likely differs from $\mathcal{H}$ in a few ways: the stratifications in $\mathcal{H}$ will probably not all have equally-sized strata, we might not be using simple random sampling to select s, and whatever distribution is chosen for Y may not be a good model for how our real response y was generated. However, we argue that the first two points do not matter, and that it is easy to choose an appropriate distribution for Y. First, our experience suggests that, in the model M, the relative size of the strata for a given $h \in \mathcal{H}_M$ makes no substantial difference in $\mathrm{E}_M[\Pr(Y_{\tilde{S}}|h)]$. This is also consistent with result (ii) of the Theorem from Section 2.3. So, it should be fine to define $\mathcal{H}_M$ with all equally-sized stratifications and use it to make decisions related to $\mathcal{H}$. Next, we argue that defining M so that $\tilde{s}$ is a simple random sample is appropriate even if the actual sampling design used to select s is, for example, stratified random sampling with respect to some $h \in \mathcal{H}$. The reason for this is that we would not want to increase the penalty term on h just because s was proportionally allocated with respect to h or decrease it because s was poorly allocated with respect to h. Selecting $\tilde{s}$ via simple random sampling represents what happens when the sample is roughly as well allocated for one stratification as for another. Finally, the multinomial type of prior distribution for y in our stepwise Bayes model means that $\Pr(y_s|h)$ only really depends on how many distinct values y_s takes, and how they are grouped with respect to h. Because we are using a stepwise Bayes approach that will tailor estimation to each possible value for unique(y_s), we can choose $\Pr(h)$ separately for each number of distinct values taken for y_s. Furthermore, the most important one of our basic baseline conditions is that the response is unrelated to the stratifications, so it's always appropriate to define M such that Y_s is unrelated to $\mathcal{H}_M$. Hence, for the prior distribution on $\mathcal{Y}_{ab}$ where $b \neq B$, we propose defining Y to be an independent vector of random variables with a uniform distribution on $\{1, 2, \ldots, a\}$ (recall that a is the number of unique values in y, and, in practice, will be the number of unique values in y_s); and for the prior distribution $\mathcal{Y}_{ab}$ where $b = B$, we propose defining Y to just be the vector $(1, 2, \ldots, N)$ (in $\Pr(Y_{\tilde{s}}|h)$, this is indistinguishable from any other distribution where observing the same value more than once occurs with

probability zero).

To get a concrete idea of what $\mathrm{E}_M[\Pr(Y_{\tilde{s}}|h)]$ might look like, let $N = 1,000$ and $n = 100$, and consider two distributions for Y: an independent vector of Bernoulli trials with probability of success equal to $1/2$, and a continuous distribution (i.e. $\Pr(y_i = y_{i'}) = 0$ for $i \neq i'$). Figure 2.2 plots the behavior of $\mathrm{E}_M[\Pr(Y_{\tilde{s}}|h)]$ for $k_h = 2, 3, 4, 5, 6$. We can see that, in the continuous case, $\Pr(Y_{\tilde{s}}|h)$ will be largest on average for the stratification h where $k_h = 6$, whereas, in the binary case, $\Pr(Y_{\tilde{s}}|h)$ will be largest on average for the stratification h where $k_h = 2$. This example reinforces the idea that we must account for a varying k_h.

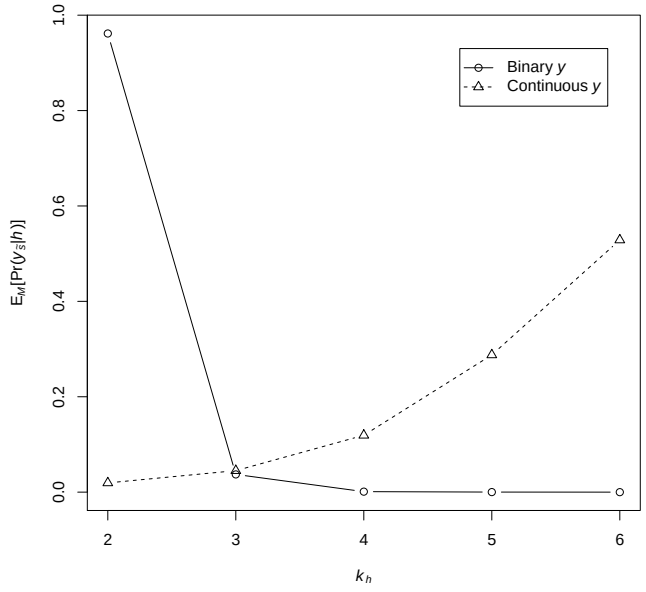

Figure 2.2: Plot of $\mathrm{E}_M[\Pr(Y_{\tilde{s}}|h)]$ onto k_h for two example distributions of Y when $N = 1000$ and $n = 100$.

One last point we should make is that we can still reflect preferences for certain $h \in \mathcal{H}$ when using this strategy to account for a varying number of strata. Just define $\Pr(h)$ as $\Pr(h|k_h)\Pr(k_h)$, choose $\Pr(k_h)$ to be inversely proportional to relevant expected value from M, and then represent any preferences in the conditional probability $\Pr(h|k_h)$.

2.5 Selecting a sample

A well-known result from Basu (1969) shows that, once s has been selected, Bayesian estimation does not depend on the sampling design. Accordingly, our estimator is appropriate for use regardless of the sampling design as long as the model producing it matches our beliefs about the population and the sample. For example, it would not be appropriate if cluster sampling was used to select s and we believed that y had a substantial intra-cluster correlation. For another example, it would be appropriate under simple random sampling or stratified random sampling using any stratification in $\mathcal{H}$. Beyond that, we can go a step farther and try to determine what type of samples will yield the most precise estimators, and can be thought of as optimal in some way.[4]

In this section, we propose a mechanism, "multi-balanced sampling", that (i) will select a sample for which multiple stratification estimation will be relatively precise, compared to other samples of the same size from the same population, and (ii) is objective. To achieve (i), we desire a mechanism which only selects samples that are "well-balanced" with respect to the full set of possible stratifications, $\mathcal{H}$. To achieve (ii), we desire a mechanism whose only restrictions from complete randomness are a fixed size and being well-balanced with respect to $\mathcal{H}$.

The multiple stratification estimation methods we have presented, in a way, generalize the typical design-based approach to stratification exemplified by stratified random sampling and post stratification. So, it is natural to consider a sampling mechanism that analogously generalizes the proportionally allocated stratified random sampling design[5]

A simple way to do this would be to draw a proportionally allocated stratified random sample based on a "refined" stratification where each unique combination of stratum memberships from all $h \in \mathcal{H}$ constitutes a stratum. This refined stratification has the property that any other stratification in $\mathcal{H}$ can be achieved by collapsing strata. Hence,

[4] We take a relatively informal, common-sense approach to determining what constitutes a good selection of s. One way of being more theoretical about picking a good sample is to use the concept of "uniform admissibility". See Mazloum and Meeden (1987) for an example.

[5] As described in the Section 2.3 footnote, we think of proportional as a good naive allocation. Also like Section 2.3, the methods described here could be adjusted to suit some known Neyman-optimal allocation if available.

such a sample would have a well-balanced allocation with respect to every stratification in $\mathcal{H}$, and would be randomly selected otherwise. When possible, we recommend this strategy. However, depending on the number of stratifications in $\mathcal{H}$, their relation to each other, and the sample size n, the refined stratification may contain too many small strata to achieve anything close to proportional allocation, even if it is possible to achieve something close to proportional allocation for each stratification individually. For example, suppose that we had five stratifications that each defined five strata. If each combination existed in the population, we would have 3,125 refined strata to sample, which may be more than the desired sample size. Of course, many of these refined strata would be non-existent or would be small could be collapsed without much concern. However, the task of collapsing refined strata so that a proportionally allocated sample from them was well-balanced with respect to each individual stratification could still be quite difficult, even though it seems that this type of sample allocation (well-balanced with respect to each of five stratifications) should be possible.

To make our proposed mechanism easy to understand, we now describe proportionally allocated stratified random sampling in an unconventional way that is more flexible with respect to determining the exact allocation of units and highlights its connection to our mechanism. Suppose a sample of size n is to be drawn from a set of N units, and that we use stratification h_0 which partitions the population into k_0 strata with N_j units in each, for $j = 1, 2, \ldots, k_0$. Then, the mechanism randomly selects (with equal probability) an element s from the set of samples that satisfy the following requirements:

- s has size n.
- $n_{jh_0} \geq 2$ for $j = 1, 2, \ldots, k_0$.
- s minimizes the objective function

$$L_0(s, h_0) = \sum_{j=1}^{k_0} w_j \left| \frac{n_{jh_0}}{n} - \frac{N_{jh_0}}{N} \right|$$

where $w_1, w_2, \ldots, w_{k_0}$ are positive real numbers that allow the allocation in some strata to matter more than others. Note that this sampling mechanism explicitly recognizes that exact proportional allocation is typically not possible. In other words, the minimum value of $L_0(s, h_0)$ taken across possible choices of s is often not zero. Now, we can easily

reframe this version of proportionally allocated stratified random sampling to apply when multiple stratifications are possible.

Suppose a sample of size n is to be drawn from $\mathcal{U}$, $\mathcal{H}$ is our set of possible stratifications with a known prior distribution $\Pr(h)$. Then, we wish to randomly select s from the set of samples that satisfy the following requirements:

mb1 s has size n.

mb2 $n_{jh} \geq 2$ for each $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$.

mb3 The objective function $L(s, \mathcal{H})$ is minimized where

$$L(s, \mathcal{H}) = \sum_{h \in \mathcal{H}} \Pr(h) \sum_{j=1}^{k_h} w_{jh} \left| \frac{n_{jh}}{n} - \frac{N_{jh}}{N} \right|$$

where w_{jh} is a positive real number for $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$. It can be easily seen that, when $\Pr(h_0) = 1$, we get the special case of proportional allocation for stratified random sampling with respect to the stratification h_0.

Note that, using our multi-balanced sampling mechanism, it is possible that some s^* is the only possible sample (i.e. it is a unique minimizer of $L(\bullet, \mathcal{H})$ among samples that satisfy the first two requirements above). In this case, the sampling mechanism contains no randomness. This may not be a problem because the (stepwise) Bayesian approach does not depend on using a random sampling design, and only objective information was used to deterministically select s^*. However, the statistician may still wish to randomly select a sample from many possible options for other reasons. This can be accommodated by replacing the requirement **mb3** with the requirement **mb3'** that $L(s, \mathcal{H}) \leq l_0$ where l_0 is chosen to balance the number and representativeness of the eligible s. For example, we can examine samples corresponding to specific quantiles of the distribution function induced on $L(s, \mathcal{H})$ by drawing s using simple random sampling restricted by requirements **mb1** and **mb2**, and then l_0 can be set to a suitable quantile.

In Chapter 4, we provide a function that implements the multi-balanced sampling mechanism in the R statistical computing language (R Core Team, 2013).

2.6 Producing sampling weights

Sampling weights are often an important ingredient in design-based inference. Many design-based point estimators, the most important example being the Horvitz-Thompson estimator, can be written as functions of the observed values and sampling weights (Särndal et al., 1992). At the same time, sampling weights create an awkward dissonance between theory and practice in design-based survey sampling; see Gelman (2007) for an extended discussion. In the most straightforward cases, a sampling weight for each unit in s is equal to the inverse of the unit's inclusion probability, as determined by the sampling design. When sampling weights are defined this way, theoretical properties are relatively easy to study, and the statistician can often show that an estimator is (almost) unbiased with respect to the sampling design or "design-consistent". However, in practice, sampling weights for large, complex surveys are usually not equal to inverse inclusion probabilities. Instead, they are based on the sampling design and other known information about the population that was not represented in the design. For example, an initial set of weights may be adjusted to reflect non-response and then adjusted again for consistency with known totals of subgroups of the population. When multiple sets of adjustments are conducted serially like this, it can become difficult to understand the design-based properties of estimation (Slud and Thibaudeau, 2010). On the other hand, failing to account for relevant known information during estimation (by adjusting sampling weights) just because this information is not represented in the sampling design seems irresponsible, putting the statistician in a double-bind. The use of a model-based approach can provide an alternative to this problem, although it is not always clear if and how information about the design should be incorporated into model-based estimation.

Strief and Meeden (2013) use a non-informative Bayesian framework to to give an alternative definition of sampling weights and to conduct estimation based on a set of sampling weights. For convenience, assume that all sampled units have distinct values. Then, to produce sampling weights, we can simply define the weight of some unit $i \in s$ to be the posterior expectation of the number of units in the population that have the same value, i.e. a the weight of a unit i

$$w_i = \mathrm{E}\left[\sum_{i'=1}^{N} \mathbb{I}(y_{i'} = y_i)|y_s\right]$$

In other words, weights can be thought of as a way of describing the composition of the population in terms of the sampled units: w_i is the number of units in the population represented by i. This is often how design-based weights are interpreted, although the justification of this interpretation is clearer here than in the design-based approach. Note that w_i is a function of y_s and model, not the design. This allows w_i to reflect known information about the population whether or not it was represented in the sampling design.

This definition of sampling weights allows us to calculate weights for a multiple stratification problem. Under the posterior distributions from Section 2.1 and Section 2.2,

$$\begin{aligned}
w_i &= \mathrm{E}\left[\sum_{i'=1}^{N} \mathbb{I}(y_{i'} = y_i)|y_s\right] \\
&= 1 + \mathrm{E}\left[\sum_{i' \in s'} \mathbb{I}(y_{i'} = y_i)|y_s\right] \\
&= 1 + \sum_{h \in \mathcal{H}} \Pr(h|y_s) \sum_{i' \in s'_{h_i h}} \mathrm{E}[\mathbb{I}(y_{i'} = y_i)|h, y_s] \\
&= 1 + \sum_{h \in \mathcal{H}} \Pr(h|y_s)(N_{h_i h} - n_{h_i h}) \frac{\epsilon_{h_i h} + 1}{r\epsilon_{h_i h} + n_{h_i h}} \\
&= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \frac{N_{h_i h} + \epsilon_{h_i h}(N_{h_i h} - n_{h_i h}) + r\epsilon_{h_i h}}{r\epsilon_{h_i h} + n_{h_i h}}
\end{aligned}$$

For a small $\epsilon_{h_i h}$, w_i is approximately equal to $\sum_{h \in \mathcal{H}} \Pr(h|y_s) N_{h_i h}/n_{h_i h}$, which is a weighted average of the design-based sampling weights produced by stratification with each $h \in \mathcal{H}$.

The estimation method presented by Strief and Meeden (2013) is based on using a set of weights w_s and the corresponding observed values y_s to produce the "Weighted Dirichlet Posterior" (WDP), which is defined as

$$\Pr(y_{s'}|w_s, y_s) = \frac{1}{N} \prod_{i \in s} \Gamma\left(\frac{n}{N} w_i + y_{y_i s'}\right)$$

The WDP derives its name from the fact that, in a complete population drawn from the posterior distribution $y_{s'}|y_s$, the proportion of units equal to each of the observed values in y_s follows a Dirichlet distribution. One reason that using the WDP for inference is attractive is that the posterior distribution can be written to contain no information pertaining to individual units besides the unique set of observed sample values, i.e. unique(y_s). This may be an important requirement in scenarios where data or an analysis are made public. We provide a simulated example of this estimation method in Section 2.7.

2.7 Simulated examples

In this section, we discuss the average performance of multiple stratification estimation relative to some typically design-based estimators as observed upon repeated sampling of three artificial populations and one "real data" population. We first discuss the artifical population examples together, and then the real data population last. For each artifical population, $N = 1600$ and y was generated according to some parametric distribution $F_\theta(y)$ where the parameter was constant within the strata defined by the "correct" stratification, denoted by h^1. In particular, y was generated in each case by the following:

Population 1:

$$y_i|h_i^1 = j \overset{\text{iid}}{\sim} \text{Normal}(\theta_j, 1)$$
$$\theta_1 = 0$$
$$\theta_2 = 1$$
$$\theta_3 = 2$$
$$\theta_4 = 3$$

Population 2:

$$y \in \{0,1,2,3,4\}^N$$
$$y_i|h_i^1 = j \overset{\text{iid}}{\sim} \text{Multinomial}(1,\theta_j)$$
$$\theta_1 = (.6,.1,.1,.1,.1)$$
$$\theta_2 = (.1,.6,.1,.1,.1)$$
$$\theta_3 = (.1,.1,.6,.1,.1)$$
$$\theta_4 = (.1,.1,.1,.1,.6)$$

Population 3:

$$y_i|h_i^1 = j \overset{\text{iid}}{\sim} \text{Poisson}(\theta_j)$$
$$\theta_1 = 1$$
$$\theta_2 = 3$$
$$\theta_3 = 5$$
$$\theta_4 = 7$$

Two stratifications in addition to h^1 were created, essentially by distorting h^1. Specifically, the three stratifications were constructed as follows:

$$h_i^1 = \begin{cases} 1, i = 1,2,\ldots,399,400 \\ 2, i = 401,402,\ldots,799,800 \\ 3, i = 801,802,\ldots,1199,1200 \\ 4, i = 1201,1202,\ldots,1599,1600 \end{cases}$$

$$h_i^2 = \begin{cases} 1, i = 1,3,\ldots,797,799 \\ 2, i = 2,4,\ldots,798,800 \\ 3, i = 801,803,\ldots,1597,1599 \\ 4, i = 802,804,\ldots,1598,1600 \end{cases}$$

$$h_i^3 = \begin{cases} 1, i = 1,3,\ldots,397,399,1201,1203,\ldots,1597,1599 \\ 2, i = 401,403,\ldots,1197,1199 \\ 3, i = 402,404,\ldots,1198,1200 \\ 4, i = 2,4,\ldots,398,400,1202,1204,\ldots,1598,1600 \end{cases}$$

Figures 2.3, 2.4, and 2.5 present graphical depictions of each population and relationship between y and h^1. We also provide R scripts that generate each population and the three stratifications in Section 2.7 (R Core Team, 2013).

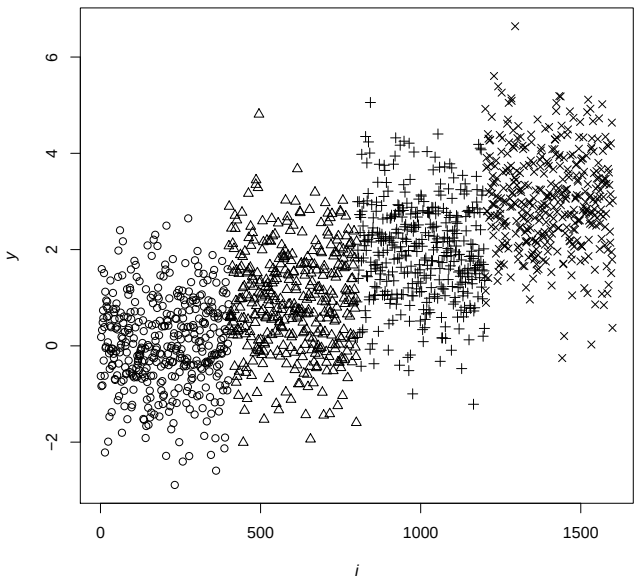

Figure 2.3: Plot of Population 1's y_i for $i \in \mathcal{U}$. The plot characters are a one-to-one function of h^1.

For each population, we drew 10,000 simple random samples of size $n = 80$. For each sample, five methods of estimation of μ were calculated. The label SRS represents the simple random sampling estimator, $\bar{y}_s$. We base the interval estimator on the Student's t distribution. Let $t^*_{\alpha,df}$ be the α^{th} quantile from the t distribution with df degrees of freedom, and let

$$v_s = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y}_s)^2$$

be the sample variance of y_s. Then, the interval estimator used for SRS was

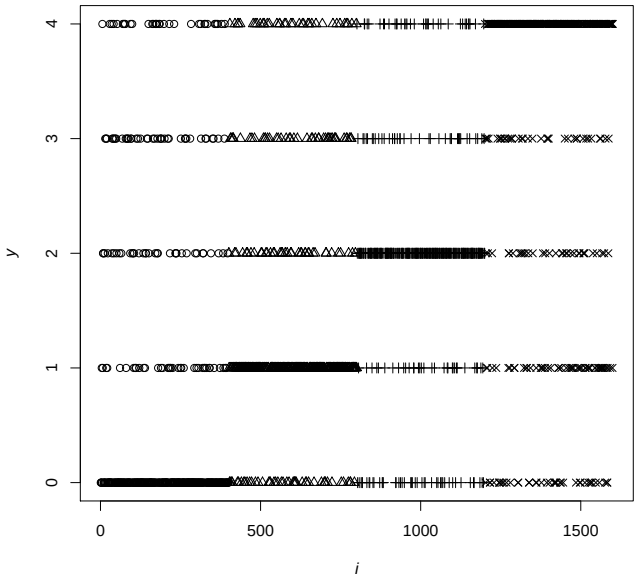

Figure 2.4: Plot of Population 2's y_i for $i \in \mathcal{U}$. The plot characters are a one-to-one function of h^1. Some "jittering" was used on the vertical axis to make the plot characters more legible.

$$\bar{y}_s \pm t^*_{.975,n-1}\sqrt{\left(\frac{1-n/N}{n}\right)v_s)}$$

This estimator should have approximately 95% coverage probability if $\bar{y}_s$ is approximately Normally distributed (which is guaranteed to hold as $n \to \infty$ by the Central Limit Theorem).

The label PS_a represents the post-stratified estimator of μ using h^a, for $a = 1, 2, 3$. We also use the label PS_r to represent the post-stratification estimator based on a "combined" or "refined" stratification. That is, let h^r be a stratification where each unique combination of stratum membership according to h^1, h^2, and h^3 defines a different h^r stratum. In other words, units i_1 and i_2 are in the same h^r stratum only if $h^a_{i_1} = h^a_{i_2}$ for $a = 1, 2, 3$. Now, for $h = h^1, h^2, h^3, h^r$, the point estimator is

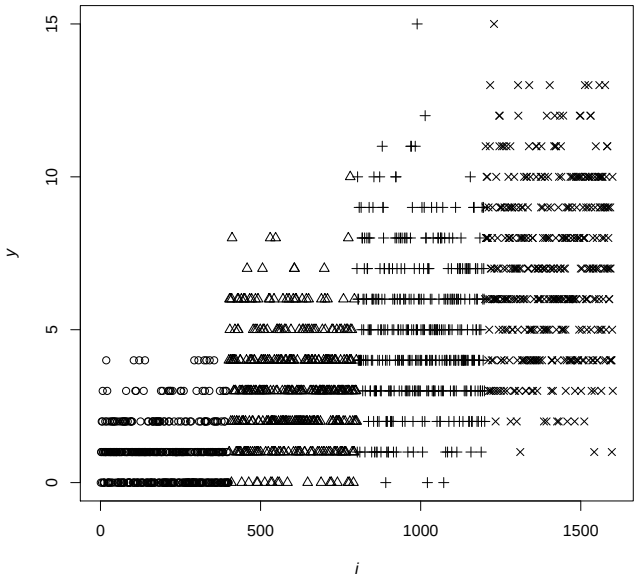

Figure 2.5: Plot of Population 3's y_i for $i \in \mathcal{U}$. The plot characters are a one-to-one function of h^1.

$$\hat{\mu}_h = \sum_{j=1}^{k_h} \frac{N_{jh}}{N} \bar{y}_{s_{jh}}$$

Then, let the sample variance of $y_{s_{jh}}$ be

$$v_{s_{jh}} = \frac{1}{n_{jh} - 1} \sum_{i \in s_{jh}} (y_i - \bar{y}^2_{s_{jh}}$$

and the interval estimator is

$$\hat{\mu}_h \pm t^*_{.975,n_e} \sqrt{\sum_{j=1}^{k_h} (N_{jh}/N)^2 \left(\frac{1 - n_{jh}/N_{jh}}{n_{jh}} \right) v_{s_{jh}}}$$

where n_e is the Satterthwaite approximation of the degrees of freedom as given by Cochran (1977, Section 5.4). Similarly to the SRS interval estimator, the PS_1, PS_2,

PS_3, and PS_r interval estimators will have approximately 95% coverage probability when $\bar{y}_{s_{jh}}$ are approximately Normally distributed for $j = 1, 2, \ldots, k_h$ and $h = h^1, h^2, h^3, h^r$, respectively.

The label MS represents estimation using the multiple stratification model presented in Section 2.2. The hyperparameters in the model were chosen as recommended, specifically $\epsilon_{jh} \propto N_{jh}$ and $\sum_{j=1}^{k_h} \epsilon_{jh} = 1/10$. For the populations considered here, this implies that $\epsilon_{jh} = 1/40$ for $j = 1, 2, 3, 4$, and $h = h^1, h^2, h^3$. We also chose $\Pr(h) = 1/3$ for $h = h^1, h^2, h^3$. The actual estimator, for each sample s drawn from the population, was a Monte Carlo approximation of $\mathrm{E}[\mu|y_s]$ calculated by drawing 1,000 independently drawn samples from the posterior distribution of $y|y_s$ and calculating μ for each. The point estimator was set to the mean of these, and the lower and upper limits for the interval estimators were set to the 0.025 and 0.975 percentiles, respectively.

Finally, the label WDP represents the WDP approach (Strief & Meeden 2013) to estimation based on the multiple stratification model as described in Section 2.6. The prescribed weights from that section were approximated as follows:

$$\begin{aligned} w_i &= \sum_{h \in \mathcal{H}} \Pr(h|y_s) \frac{N_{h_i h} + \epsilon_{h_i h}(N_{h_i h} - n_{h_i h}) + r\epsilon_{h_i h}}{r\epsilon_{h_i h} + n_{h_i h}} \\ &\approx \sum_{h \in \mathcal{H}} \Pr(h|y_s) \frac{N_{h_i h}}{n_{h_i h}} \end{aligned}$$

where $\Pr(h|y_s)$ was calculated using the same ϵ_{jh} hyperparameters as for the MS estimator. Once again, the actual estimator for each s drawn from the population was a Monte Carlo approximation of $\mathrm{E}[\mu|y_s]$ using 1,000 independently drawn samples from the posterior from the WDP distribution for $y|y_s$, calculating μ for each, and using the mean for the point estimator and 0.025 and 0.975 percentiles for lower and upper limits of the interval estimator, respectively.

The simulation results for Populations 1, 2, and 3 are given in Table 2.7. Before discussing the results for each population in turn, we will provide a couple of general comments. First, we can see that, for each population, the observed bias upon repeatedly sampling is approximately zero for each estimator. Hence, differences in observed performance - which we measure by mean absolute error - are essentially differences in

the variances of estimators across samples. Second, the performance of the PS_1 and SRS estimators, can be used as baselines with which to compare the performance of the MS estimator on each population; the PS_1 estimator can be thought of as an upper bound in terms of performance because it is based on knowing the correct stratification, and the SRS estimator can be thought of as a lower bound because it eschews any stratification information. Third, the PS_r estimator is seen to be a fine alternative to the multiple stratification estimation approach when there is uncertainty regarding the correct stratification and the sample sizes in each stratum of the refined stratification are sufficiently large. In the stratifications h^1, h^2, and h^3 generated for these examples, the population-level sizes of the strata defined by h^r were all fairly large, so the corresponding sample sizes based on a simple random sample would typically be large enough. In other cases, this may not be true, necessitating collapsing some of the strata defined by h^r if it is to be used. Fourth, as expected, the WDP point estimator performed very similarly to the MS estimator on all three populations, suggesting that it is a fine alternative. The drawback of using this technique, judging from these examples, is the inflated coverage probability of its interval estimator. Finally, recall that, since the hyperparameters were chosen to be small ($\epsilon_{jh} = 1/40$), the MS estimator is approximately equal to a weighted average of PS_1, PS_2, PS_3. Hence, the relative performance MS compared to the post-stratification estimators mostly depends on the behavior of the mixture weights, $\Pr(h|y_s)$.

Population 1, where y consists of all distinct values, represents a baseline case. For a sample y_s drawn from this population, any preference for a particular stratification in $\Pr(h|y_s)$ will be based on the sample allocation of s with respect h because $\Pr(h)$ is a constant, and there is always zero within-stratum homogeneity for each h. Also, because simple random sampling is used to select s on each draw and h^1, h^2, h^3 are just permutations of each other, the sample allocation of y_s with respect to h (which can be thought of as a random variable under simple random sampling) will exhibit the same behavior for $h = h^1, h^2, h^3$. Consequently, upon repeated sampling, the average value of $\Pr(h|y_s)$ should be 1/3 for $h = h^1, h^2, h^3$. Therefore, it is not surprising that the mean absolute error of MS (0.111) is approximately equal to the average of the mean absolute errors for the post-stratification estimators (0.109). We can also see that the coverage probability of each interval estimator is approximately equal to its target value

of 95%. Given that the (stratum-) sample means used to calculate the SRS and post-stratification estimators are approximately Normally distributed, this is theoretically guaranteed for those estimators, but not for the MS estimator. Finally, the mean width of the interval estimators is negatively correlated with the mean absolute errors of the point estimators.

In contrast to Population 1, the highly discrete response of Population 2 represents an ideal scenario for the MS estimator. Here, the within-stratum homogeneity of y_s will, on average, be larger for h^1 than h^2 or h^3. In other words, the actual values observed in y_s should help $\Pr(h|y_s)$ select the correct the stratification. The observed results are consistent with this argument: MS and PS_1 have approximately the same mean absolute error. Hence, in an ideal scenario, we may not be sacrificing any performance by not knowing the correct stratification. For this population, the coverage probabilities for each interval is still approximately equal to the target value of 95%, but the MS coverage probability is slightly lower at 93.2%. As mentioned above, the MS interval estimator is the only one that is not based on a Central Limit Theorem, and has no theoretical guarantee related to its coverage probability. On the other hand, this mild under-coverage may be made up for by its smaller mean width.

Population 3 represents an intermediate case between Populations 1 and 2, and the relative performance of the MS estimator follows suit. In this population, y does not consist of all distinct values, and there is more within-stratum homogeneity when h^1 is used than h^2 or h^3 is used. However, instead of y_i taking on one of only five distinct values for $i = 1, 2, \ldots, N$, there are 15 unique values appearing in y for Population 3. Hence, the ability of $\Pr(h|y_s)$ to detect disparate the within-stratum homogeneity is not as strong as for Population 2. Concordantly, the mean absolute error for the MS estimator is not as small as for the PS_1 estimator, but it is quite close (0.176 versus 0.179). The MS interval estimator also has a smaller mean width than other estimators on this population, and has a coverage probability (94.2%) approximately equal to the target probability of 95%.

Taken together, these three sets of simulation studies provide anecdotal evidence for a few points. First, we can expect the performance of multiple stratification estimation to be no worse than if post-stratification with the "worst" stratification in $\mathcal{H}$ had been used and no better than if post-stratification with the "best" stratification had been

used. This type of behavior should be expected from a model-averaging approach. Second, the ability of the mixture weights $\Pr(h|y_s)$ to help select a good stratification depends on the nature of the population: within-stratum homogeneity as defined by a preponderance of repeated values within a stratum must be present. This is an important limitation of the multiple stratification model constructed here, and can be seen as the price paid for avoiding parametric assumptions in a non-informative Bayesian approach. We hypothesize that, in order for $\Pr(h|y_s)$ to be able to discriminate between stratifications when y consists of all distinct values, the distribution $y_s|h$ would have to be a parametric continuous distribution, e.g. a Normal distribution. Third, the results for Population 3 are particularly encouraging. Because Population 2 essentially matches the prior distributions used in the stepwise Bayes model, if the MS estimator only performed well on examples like Population 2, the argument could be made that the multiple stratification model only works well when the prior distribution used for inference is correct. Instead, the good performance of the MS estimator on Population 2 suggests that multiple stratification estimation should work well on any population where y is "discrete enough" so that repeated values are regularly observed.

Finally, we discuss the real data population example. This example is not intended to provide a practical application, but rather evidence that our multiple stratification model can be useful on populations that were not artificially generated. For this example, we obtained 2011 American Community Survey (ACS) data, which is collected by the U.S. Census Bureau, from the Integrated Public Use Microdata Series (IPUMS-USA) (Ruggles et al., 2010). Specifically, we looked at three variables for all 2011 ACS respondents: age, total income from the previous year, and education. We treated education as the response and age and income as auxiliary variables on which to base a stratification. After removing missing data, the population consisted of $N = 17,238$ respondents. In order to define y, we collapsed a few categories in the the education variable. For the $i = 1, 2, \ldots, N$,

Real Data Population:

$$y_i = \begin{cases} 1, & \text{if } i\text{'s response was N/A or no schooling} \\ 2, & \text{if } i\text{'s response was nursery school to grade 4} \\ 3, & \text{if } i\text{'s response was grade 5, 6, 7, or 8} \\ 4, & \text{if } i\text{'s response was grade 9, 10, or 11} \\ 5, & \text{if } i\text{'s response was grade 12} \\ 6, & \text{if } i\text{'s response was 1 year of college} \\ 7, & \text{if } i\text{'s response was 2 years of college} \\ 8, & \text{if } i\text{'s response was 4 years of college} \\ 9, & \text{if } i\text{'s response was 5 or more years of college} \end{cases}$$

Using the age and income data, we created the stratification h^1. For the $i = 1, 2, \ldots, N$,

$$h_i = \begin{cases} 1, & \text{if } i \text{ was below the age and income medians} \\ 2, & \text{if } i \text{ was below the age median and above the income median} \\ 3, & \text{if } i \text{ was above the age median and below income age median} \\ 4, & \text{if } i \text{ was above the age and income medians} \end{cases}$$

Table 2.7 describes the composition of y in relation to h^1.

In addition to h^1, we created three more stratifications. One of them, h^2, was a "distorted" version of h^1. First, h^2 was set equal to h^2. Then, $4,000$ units were randomly selected and had their values changed to either 5 or 6 ($2,000$ of each). The other two stratifications were artificially generated: h^3 from a uniform distribution on $\{1, 2, 3\}$ and h^4 from a uniform distribution on $\{1, 2, 3, 4, 5\}$. The idea behind this selection of four stratifications is that stratification on h^1 will produce the greatest precision, stratification on h^2 will produce an improvement relative to using the sample mean but not as much as h^1, and h^3 and h^4 will produce no gains relative to using the sample mean. Furthermore, a varying k_h has to be dealt with. If preference is given to a small k_h, substantial influence may be given to h^3; and if preference is given to a large k_h, substantial influence may be given to h^4. On the other hand, if we are able to effectively handle the varying k_h in this problem, the multiple stratification estimator should perform as well as stratification on h^1.

The estimators used for this population were SRS, PS_1, PS_2, PS_3, PS_4, and MS. The point estimation and interval estimation methods are all the same as they described earlier with one exception. The MS method used the approach described in Section 2.4 to account for a varying k_h. Specifically, the model M was used where $\mathcal{H}_M$ consisted of stratifications with $k_h \in \{3, 4, 5, 6\}$ and equally-sized strata, and Y was generated with the discrete uniform distribution on unique(y_s).

The results for this population (Table 2.7) are based on drawing $10,000$ samples of $n = 500$, and estimating the population using each of the described methods. The results show that PS_1 and PS_2 achieved about a 10% and 8% reduction in mean absolute error relative to SRS, respectively, while PS_3 and PS_4 performed about the same as SRS. The MS estimator was able to approximately match the PS_1 estimator in terms of mean absolute error. This suggests that the posterior distribution $\Pr(h|y_s)$ was able to deal with the varying k_h, and base inference almost entirely on the best available stratification. This simulated example provides more evidence that, when there are multiple possible stratifications and the response is discrete, our stepwise Bayes model allows the statistician to do about as well on average as if the correct stratification was known.

Artifical population generation

Here we provide scripts that can be used to generate the artificial populations and stratifications discussed above in R Core Team (2013). The real data population and its stratifications are not provided here.

Population 1:

```
set.seed(1)


## y
stratum.size <- 400
theta <- c(rep(0,stratum.size),
           rep(1,stratum.size),
           rep(2,stratum.size),
           rep(3,stratum.size))
```

```
f.gen.y <- function(u) rnorm(1,u)
y.pop <- sapply(theta, f.gen.y)
N <- length(y.pop)

## stratifications
h.perfect <- as.integer(factor(theta))
h.good <- c(rep(1:2,stratum.size),
            rep(3:4,stratum.size))
h.bad <- c(rep(c(1,4),stratum.size/2),
           rep(2:3,stratum.size),
           rep(c(1,4),stratum.size/2))
H <- data.frame(h.perfect,h.good,h.bad)
```

Population 2:

```
set.seed(1)

## y
stratum.size <- 400
theta <- c(rep(0,stratum.size),
           rep(1,stratum.size),
           rep(2,stratum.size),
           rep(3,stratum.size))
f.theta <- function(u){
    if(u==0) out <- c(6,1,1,1,1)
    if(u==1) out <- c(1,6,1,1,1)
    if(u==2) out <- c(1,1,6,1,1)
    if(u==3) out <- c(1,1,1,1,6)
    return(out)
}
f.gen.y <- function(u) sample(0:4,size=1,prob=f.theta(u))
y.pop <- sapply(theta, f.gen.y)
```

```
N <- length(y.pop)

## stratifications
h.perfect <- theta+1
h.good <- c(rep(1:2,stratum.size),
            rep(3:4,stratum.size))
h.bad <- c(rep(c(1,4),stratum.size/2),
           rep(2:3,stratum.size),
           rep(c(1,4),stratum.size/2))
H <- data.frame(h.perfect,h.good,h.bad)
\begin{verbatim}

\noindent \textbf{Population 3:}
\begin{verbatim}
set.seed(1)

## y
stratum.size <- 400
lambda <- c(rep(1,stratum.size),
            rep(3,stratum.size),
            rep(5,stratum.size),
            rep(7,stratum.size))

f.gen.y <- function(u) rpois(1,u)
y.pop <- sapply(lambda, f.gen.y)
N <- length(y.pop)

## stratifications
h.perfect <- as.integer(factor(lambda))
h.good <- c(rep(1:2,stratum.size),
            rep(3:4,stratum.size))
h.bad <- c(rep(c(1,4),stratum.size/2),
```

```
            rep(2:3,stratum.size),
            rep(c(1,4),stratum.size/2))
H <- data.frame(h.perfect,h.good,h.bad)
```

Table 2.1: Simulation results for the artificial populations: *mae* gives the mean absolute error of the point estimator, *bias* gives the bias of the point estimator, *lower* gives the mean of the lower interval estimator limit, *width* gives the mean width of the interval estimator, and *cover* gives the coverage probability of the interval estimator.

Population 1	*mae*	*bias*	*lower*	*width*	*cover*
PS_1	0.092	0.001	1.263	0.456	0.947
PS_r	0.095	0.001	1.252	0.479	0.947
PS_2	0.102	0.001	1.239	0.505	0.947
WDP	0.109	0.000	1.172	0.639	0.978
MS	0.111	0.000	1.222	0.539	0.946
SRS	0.132	0.000	1.161	0.658	0.953
PS_3	0.135	0.000	1.152	0.676	0.952
Population 2	*mae*	*bias*	*lower*	*width*	*cover*
MS	0.111	0.000	1.577	0.511	0.932
PS_1	0.112	-0.001	1.553	0.554	0.950
WDP	0.112	-0.001	1.528	0.610	0.970
PS_r	0.115	-0.001	1.538	0.583	0.950
PS_2	0.117	-0.001	1.539	0.581	0.950
SRS	0.129	-0.002	1.514	0.630	0.946
PS_3	0.131	-0.002	1.506	0.646	0.949
Population 3	*mae*	*bias*	*lower*	*width*	*cover*
PS_1	0.180	0.000	3.511	0.911	0.954
WDP	0.181	0.000	3.359	1.260	0.993
MS	0.183	0.028	3.572	0.867	0.940
PS_r	0.186	0.001	3.486	0.963	0.954
PS_2	0.203	0.000	3.458	1.017	0.956
SRS	0.259	0.001	3.319	1.297	0.950
PS_3	0.265	0.001	3.300	1.334	0.950

Table 2.2: Composition of y and h^1 for the real data population.

y	$h^1 = 1$	$h^1 = 2$	$h^1 = 3$	$h^1 = 4$
1	64	7	135	16
2	30	3	74	8
3	278	30	316	64
4	1200	64	394	125
5	2056	927	2075	1374
6	970	499	541	597
7	313	361	258	329
8	478	952	381	853
9	117	518	148	683

Table 2.3: Simulation results for the real data population: *mae* gives the mean absolute error of the point estimator, *bias* gives the bias of the point estimator, *lower* gives the mean of the lower interval estimator limit, *width* gives the mean width of the interval estimator, and *cover* gives the coverage probability of the interval estimator.

	mae	*bias*	*lower*	*width*	*cover*
PS_1	0.054	-0.000	5.709	0.275	0.956
MS	0.054	-0.001	5.709	0.273	0.955
PS_2	0.055	0.000	5.707	0.279	0.956
SRS	0.060	0.001	5.697	0.300	0.957
PS_4	0.060	0.001	5.697	0.301	0.957
PS_3	0.060	0.001	5.697	0.301	0.954

Chapter 3

Multiple Stratification for a Non-Response Problem

In this chapter, we present an alternative version of our multiple stratification model that can be used for a type of non-response problem discussed by Hansen and Hurwitz (1946). In that paper, the authors describe a scenario where an initial sampling attempt using mail questionnaires results in some non-response, and a subsample of these non-respondents is taken using field interviews. Estimation is then conducted using double-sampling framework, also known as two-phase sampling. The basic idea of the strategy in Hansen and Hurwitz (1946) is to use the response rate from the initial sample to estimate the proportions of the population that fall into strata defined as "would-be responders" and "would-be non-responders", and then to conduct estimation by post-stratifying with these estimated stratum weights. We incorporate the multiple stratification approach into this problem by supposing that there is some stratification, h, such that "response behavior" is more homogenous within the strata defined by h. Hence, if we knew what h was, we could use it to improve our estimates of the proportions of "would-be responders" and "would-be non-responders". Instead, we suppose that have a set of stratifications, $\mathcal{H}$, and show how to use multiple stratification to average estimation across them.

Because the type of non-response problem that we consider is actually a special case of double-sampling, we develop the models in this chapter in the broader framework

of double-sampling. We define double-sampling as the case where a stratification, r, is observed for units in a size-m subset t of $\mathcal{U}$, and the response, y, is observed for units in the size-n subset s of t, i.e. $s \subset t \subset \mathcal{U}$. We use t' and s' to refer to the complements of t and s, respectively. We assume that, for each stratum u observed in r, there are at least two units in s where r equals u. This assumption ensures that our sample of y is at least minimally representative of the population as stratified by r. We also assume that we are only interested in r to the extent that it helps us estimate parameters of y, and not for its own sake. The procedure used to select t and s are not important for our basic set-up.

This double-sampling framework is applied to non-response when r_i describes the "response behavior" of the i^{th} unit. Specifically, let $r_i = 0$ if i will respond on the first sampling attempt, and $r_i = 1$ if it responds on the second attempt. Suppose that we select t with the intention of observing y_i for all $i \in t$, but observe it for $i \in s_1$ where

$$s_1 = \{i \in t : r_i = 0\}$$

Next, we select a subsample $s_2 \subset t - s_1$, and, presumably through more expensive means, are able to observe y_i for each $i \in s_2$. In Hansen and Hurwitz (1946), this subsample is observed through field interviews; subsample observations can also be referred to as "callbacks". Then, we define s as $s_1 \cup s_2$, and the non-response problem fits into the double-sampling framework. More generally, we may still fail to observe y_i for all $i \in s_2$. If this is the case, we can let $r_i = 2$ for the units i that failed to respond on the second attempt, and repeat the subsampling procedure on the two-time non-responders. This strategy can be iterated an arbitrary number of times, each time whittling down the number of non-responders, but there are essentially two ways the situation can resolve: (1) there is no non-response in final sampling attempt, or (2) some "hard core" (excellent terminology taken from Cochran, 1977) of units sampled in the final attempt are not observed. In the case of (1), we can once again have confidence that each type of unit in the population is well-represented in our sample. In the case of (2), some assumption must be made, e.g. that the "hard core" are exchangeable with the non-responding units observed in the final sampling attempt, or the responses of the "hard core" may be modeled using auxiliary information. We discuss this issue a little further in Section 3.4, but assume that outcome (1) is achieved in the multiple

stratification models developed here.

As briefly described above, we can incorporate the multiple stratification approach into this problem in the following way. Assume that some stratification from a collection of possible stratifications, i.e. some $h \in H$, provides a good grouping of units with respect to r. Hence, after observing r_t, the stratifications in H can be help us determine what $r_{t'}$ looks like. This extra information about r will help us estimate parameters of y. Importantly, though, we assume that $y \perp h \,|r$, i.e. y and h are conditionally independent given r. That is, in the non-response setting, $\mathcal{H}$ only helps us estimate the prevalence of "would-be non-responders" in the population.

In Section 3.1, we define a fully Bayes model for the "binary" case where r and y lie in $\{0,1\}^N$ because the stepwise Bayes technique will not be necessary. Then, in Section 3.2 we define the model for the general case using the stepwise Bayes technique. In Section 3.3, we present two simulated examples of our stepwise Bayes model. Finally, in Section 3.4, we discuss potential further developments that would allow the multiple stratification approach to be incorporated into other types of non-response problems.

3.1 A binary response

Let $\mathcal{R}$ be the parameter set for r and $\mathcal{Y}$ be the parameter set for y. In this section, we assume that $\mathcal{R}$ and $\mathcal{Y}$ both equal $\{0,1\}^N$. We carry forward the notational conventions used in Chapter 2. For example, $r_{t_{jh}}$ is a set containing the values of r for units $i \in t_{jh}$, for some $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$. In addition, we use the index pair ur for creating subsets related to strata defined by a given value of r: $\mathcal{U}_{ur}$ and s_{ur} are the the sets of units $i \in \mathcal{U}$ and $i \in s$, respectively, such that $r_i = u$ for $u = 0, 1$.

Now, we present the basic structure of the prior distribution. The second line highlights the conditional independence between y and h given r, and the third line shows the independence between strata in r and y.

$$\begin{aligned}\Pr(h, r, y) &= \Pr(h)\Pr(r|h)\Pr(y|h, r) \\ &= \Pr(h)\Pr(r|h)\Pr(y|r) \\ &= \Pr(h)\prod_{j=1}^{k_h}\Pr(r_{\mathcal{U}_{jh}}|h)\prod_{u=0}^{1}\Pr(y_{\mathcal{U}_{ur}}|r)\end{aligned}$$

To get the marginal prior distribution for y, we sum over the distributions of h and r.

$$\Pr(y) = \sum_{h\in\mathcal{H}}\Pr(h)\sum_{r\in\mathcal{R}}\prod_{j=1}^{k_h}\Pr(r_{\mathcal{U}_{jh}}|h)\prod_{u=0}^{1}\Pr(y_{\mathcal{U}_{ur}}|r)$$

Now, we can see that r has exactly the same structural relationship with $\mathcal{H}$ as y had in Chapter 2. The relationship between y and r in the prior distribution is also similar to the relationship between y and h from Chapter 2: i.e. $y|r$ is a group of independent strata. There is an important difference in that we will actually observe r on some units, though.

Next, we define $\Pr(r_{\mathcal{U}_{jh}}|h)$ for $h \in \mathcal{H}$ and $j = 1, 2, \ldots, k_h$, and $\Pr(y_{\mathcal{U}_{ur}}|r)$ for $r \in \mathcal{R}$ and $u = 0, 1$.

$$\begin{aligned}\Pr(r_{\mathcal{U}_{jh}}|h) &= \frac{\Gamma(2\epsilon_{jh})\Gamma(\epsilon_{jh} + r_{1\mathcal{U}_{jh}})\Gamma(\epsilon_{jh} + r_{0\mathcal{U}_{jh}})}{\Gamma(\epsilon_{jh})^2\Gamma(2\epsilon_{jh} + N_{jh})} \\ \Pr(y_{\mathcal{U}_{ur}}|r) &= \frac{\Gamma(2\eta_u)\Gamma(\eta_u + y_{1\mathcal{U}_{ur}})\Gamma(\eta_u + y_{0\mathcal{U}_{ur}})}{\Gamma(\eta_u)^2\Gamma(2\eta_u + N_{ur})}\end{aligned}$$

Both of these distributions match $\Pr(y_{\mathcal{U}_{jh}}|h)$ from Section 2.1. The hyperparameters for $r_{\mathcal{U}_{jh}}|h$ are the same as we used in $y_{\mathcal{U}_{jh}}|h$, and we recommend choosing them in the same way for the same reasons. For $y_{\mathcal{U}_{ur}}|r$, we define a new hyperparameter $\eta = (\eta_1, \eta_0)$. Unlike the ϵ_{jh} hyperparameters, η has no real use besides producing a proper prior distribution, we do not define it to depend on on r, and it will not appear Section 3.2's stepwise Bayes model. Our recommendation is simply to choose it to be small, and we will see below how this is a good non-informative Bayesian choice for the posterior distribution momentarily. Both of these distributions also have convenient forms for the marginal probabilities associated with the samples t and s.

$$\Pr(r_{t_{jh}}|h) = \frac{\Gamma(2\epsilon_{jh})\Gamma(\epsilon_{jh} + r_{1t_{jh}})\Gamma(\epsilon_{jh} + r_{0t_{jh}})}{\Gamma(\epsilon_{jh})^2\Gamma(2\epsilon_{jh} + m_{jh})}$$
$$\Pr(y_{s_{ur}}|r) = \frac{\Gamma(2\eta_u)\Gamma(\eta_u + y_{1s_{ur}})\Gamma(\eta_u + y_{0s_{ur}})}{\Gamma(\eta_u)^2\Gamma(2\eta_u + n_{ur})}$$

where m_{jh} refers to the number of units i in t where $h_i = j$, and n_{ur} refers to the number of units i in s where $r_i = u$.

Now, we can examine the posterior distribution implied by observing r_t and y_s. The general structure is the following:

$$\begin{aligned}\Pr(y_{s'}|r_t, y_s) &= \sum_{h\in\mathcal{H}} \Pr(h|r_t) \sum_{r\in\mathcal{R}} \Pr(r_{t'}|h, r_t) \Pr(y_{s'}|r, y_s) \\ &= \sum_{h\in\mathcal{H}} \Pr(h|r_t) \sum_{r\in\mathcal{R}} \left(\prod_{j=1}^{k_h} \Pr(r_{t'_{jh}}|r_{t_{jh}}, h)\right) \left(\prod_{u=0}^{1} \Pr(y_{s'_{ur}}|y_{s_{ur}}, r)\right)\end{aligned}$$

Hence, $y|r_t, y_r$ is essentially a two-stage recursive finite mixture model: we can pick an h from $h|r_t$, pick an r from $r_{t'}|r_t, h$, and then we have our conditional model $y_{s'}|r, y_s$. An interesting difference between the posterior distribution for y here and the one from Section 2.1 is that, in Section 2.1, the "mixture weights" for the finite mixture model depended on y_s, but here they do not. In this model, the only data influencing the mixture weights, both $\Pr(h|r_t)$ and $\Pr(r|r_t, h)$, is r_t. This is the reason that our recommendation for choosing η is to just make it small. Its only appearance in the posterior distribution is in $\Pr(y_{s'}|y_s, r)$, so it does not influence the mixture weights like ϵ_{jh} does.

At this point, we can see that the posterior distribution for r is essentially identical to the posterior distribution for y seen in Section 2.1.

$$\Pr(r_{t'}|r_t) = \sum_{h\in\mathcal{H}} \Pr(h|r_t) \prod_{j=1}^{k_h} Pr(r_{t'_{jh}}|h, r_{t_{jh}})$$
$$\Pr(r_{t'_{jh}}|h, r_{t_{jh}}) = \frac{\Gamma(2\epsilon_{jh} + m_{jh})\Gamma(\epsilon_{jh} + r_{1\mathcal{U}_{jh}})\Gamma(\epsilon_{jh} + r_{0\mathcal{U}_{jh}})}{\Gamma(2\epsilon_{jh} + N_{jh})\Gamma(\epsilon_{jh} + r_{1t_{jh}})\Gamma(\epsilon_{jh} + r_{0t_{jh}})}$$
$$\Pr(h|r_t) \propto \Pr(h) \prod_{j=1}^{k_h} \Pr(r_{t_{jh}}|h)$$

Hence, we refer the reader to Section 2.1 for more details on characteristics of the posterior distribution. Next, we give the distribution $y_{s'}|y_s, r$.

$$\begin{aligned}\Pr(y_{s'}|y_s, r) &= \prod_{u=0}^{1} \Pr(y_{s'_{ur}}|y_{s_{ur}}, r) \\ &= \prod_{u=0}^{1} \frac{\Gamma(2\eta_u + n_{ur})\Gamma(\eta_u + y_{1\mathcal{U}_{ur}})\Gamma(\eta_u + y_{0\mathcal{U}_{ur}})}{\Gamma(2\eta_u + N_{ur})\Gamma(\eta_u + y_{1s_{ur}})\Gamma(\eta_u + y_{0s_{ur}})}\end{aligned}$$

Now, we will provide some examples of estimation using this posterior distribution. First, we derive the posterior expected value of an unseen unit. For $i \in s'$,

$$\mathrm{E}[y_i|y_s, r_t] = \sum_{h\in\mathcal{H}} \Pr(h|r_t) \sum_{r\in\mathcal{R}} \Pr(r|r_t, h)\mathrm{E}[y_i|y_s, r]$$

Note the features of the posterior distribution that appear here. First, $\mathrm{E}[y_i|y_s, r_t]$ has the two-stage finite mixture model representation described above. Second, the mixture weights, which depend only on r_t and the choice of ϵ_{jh}, are expected to be relatively large for $h \in \mathcal{H}$ that lead to a good sample allocation of t and within-stratum homogeneity of r_t. Third, conditional on some $h \in \mathcal{H}$, we expect $\Pr(r|r_t, h)$ to be large for $r \in \mathcal{R}$ where the proportion of units in t'_{jh} equal to 1 is close to the observed proportion in t_{jh} for $j = 1, 2, \ldots, k_h$. Now, we look at the conditional posterior expectation. Given some $r \in \mathcal{R}$ and supposing that $i \in s'_{ur}$, we can make essentially same calculation we made in Section 2.1.

$$\mathrm{E}[y_i|y_s, r] = \frac{\eta_u + y_{1s_{ur}}}{2\eta_u + n_{ur}}$$

Note that $\mathrm{E}[y_i|y_s, r]$ will approach the stratum sample mean $\bar{y}_{s_{ur}}$ as $\eta_u \to 0$.

Like before, this calculation can be used to examine the estimator, under squared-error loss, for any parameter $\gamma(y)$ that is a linear function of y.

$$\mathrm{E}[\gamma(y)|y_s, r_t] = \sum_{h\in\mathcal{H}} \Pr(h|r_t) \sum_{r\in\mathcal{R}} \Pr(r|r_t, h)\gamma(\mathrm{E}[y|y_s, r])$$

Finally, we provide the example of estimating μ.

$$\mathrm{E}[\mu|y_s, r_t] = \sum_{h\in\mathcal{H}} \Pr(h|r_t) \sum_{r\in\mathcal{R}} \Pr(r|r_t, h) \left(\frac{n_{ur}}{N_{ur}} \bar{y}_{s_{ur}} + \left(\frac{N_{ur} - n_{ur}}{N_{ur}} \right) \mathrm{E}[y_i|y_s, r, i \in s'_{ur}] \right)$$

Just as in Section 2.1, as $\eta \to (0, 0)$, this estimator will approach a weighted average of design-based stratified estimators of μ.

3.2 Stepwise Bayes model

Now, we present the multiple stratification stepwise Bayes model for a double-sampling scenario. The resulting posterior distribution will essentially be a weighted average of multiple versions of the posterior distribution seen in the Vardeman and Meeden (1984) paper. The stepwise model differs from the fully Bayes model above in two main ways. First, it allows the set of distinct values that can appear in r and y with positive probability to be determined by the sample - an analogous extension was how the fully and stepwise Bayes models from Chapter 2 differed. Second, it removes the hyperparameter η from the model. The hyperparameters ϵ_{jh} remain as they did in Section 2.2, in order to achieve a good relationship between $\Pr(h|r_t)$ and the sample allocation of t with respect to h. This model necessitates a fairly involved ordered collection of restricted parameter sets because, in order to accomplish this task, r must be incorporated into the definition of the full and restricted parameter sets. As we proceed, the astute reader will notice that the case where r and y are both binary can be handled by this model also, and will lead to a posterior distribution that only differs in that the η hyperparameter will be absent (reduced to zero).

First, we handle notation and assumptions related to r. Assume that the possible number of strata given by r is less than or equal to some positive integer C, and that the labels for these strata are simply the integers $1, 2, \ldots, C$. Furthermore, if c strata are defined by r, for some $c \in \{2, 3, \ldots, C\}$, we assume that the strata labels are $1, 2, \ldots, c$. Note that this assumption is not restrictive in any practical sense because label used to identify a stratum does not influence estimation.

Next, we handle notation and assumptions related to y. Let B once again be the set of distinct values that may appear in y, and assume that B contains A values where A is a finite number. In the last stepwise Bayes model, we used the index b to

denote a subset of B that identified unique(y). Here, given that r defines c strata, for some $c \in \{2, 3, \ldots, C\}$, we extend this notation to the list $b = (b_1, b_2, \ldots, b_c)$ where $b_u \subseteq B$ determines unique($y_{\mathcal{U}_{ur}}$) for $u = 1, 2, \ldots, c$. Also, we use the index vector $a = (a_1, a_2, \ldots, a_c)$ where a_u gives the number of elements in b_u for $u = 1, 2, \ldots, c$.

Now, we first define the ordered collection of disjoint restricted parameter sets on which our prior distributions will be defined, and then we will define the prior distribution for a generic restricted parameter set. First, we define the full parameter set as

$$\Omega = \mathcal{R} \times \mathcal{Y}$$

where

$$\mathcal{R} = \{r \in \{1, 2, \ldots, C\}^N : \text{unique}(r) = \{1, 2, \ldots, c\} \text{ for some } c \in \{2, 3, \ldots, C\}\}$$
$$\mathcal{Y} = B^N$$

The definition of $\mathcal{R}$ may seem slightly strange, but it is just the definition that follows from assuming that the integers $1, 2, \ldots, c$ are used as strata labels when r defines c strata, for $c \in \{2, 3, \ldots, C\}$.

Next, a generic restricted parameter set is defined as

$$\Omega_{cab} = \{(r, y) \in \Omega : r \in \{1, 2, \ldots, c\}^N, \text{unique}(y_{\mathcal{U}_{ur}}) = b_u \text{ for } u = 1, 2, \ldots, c\}$$

Similarly to the last stepwise Bayes model, this set requires $y_{\mathcal{U}_{ur}}$ to take on all values in b_u at least once for $u = 1, 2, \ldots, c$. Note that the same $r \in \mathcal{R}$ or the same $y \in \mathcal{Y}$ may appear in multiple restricted sets, but every pair $(r, y) \in \Omega$ will appear in exactly one restricted set

Now we can define the ordering of the restricted parameter sets along the index *cab*. The ordering runs through a series of nested loops. For each step in loop one, loop two runs through each of its steps in order; for each step in loop two, loop three runs through each of its steps, and so on. The outer most loop is the index c, which goes from 2 to C. For a given value of c, we proceed through all possible values of (a, b), but this process is

divided into c different loops, where the u^{th} loop corresponds to the index pair (a_u, b_u), for $u = 1, 2, \ldots, c$. The nesting of these c loops does not matter, but it is fine to let (a_{u+1}, b_{u+1}) be nested below (a_u, b_u), for $u = 1, 2, \ldots, c-1$. That is, (a_c, b_c) proceeds in order through all possible steps while holding $((a_1, b_1), (a_2, b_2), \ldots, (a_{c-1}, b_{c-1})$ constant. Then, (a_{c-1}, b_{c-1}) is moved to its next value, and (a_c, b_c) runs through its cycle again. When (a_{c-1}, b_{c-1}) has gone through its entire cycle, with (a_c, b_c) completing its cycle at each step, (a_{c-2}, b_{c-2}) is moved to its next value, and so on. Now, the ordering through which (a_u, b_u) cycles, for $u = 1, 2, \ldots, c$ is the same phase/step set-up we used in the last stepwise Bayes model. Specifically, a_u proceeds from 1 to A, and, for each value of a_u, b_u cycles through all size a_u subsets of B. As before, the order of the "phases" through which a_u proceeds matters, but the order of the "steps" through which b_u proceeds does not. One way to think about this ordering is that, for each c, we progress through the same ordering as used for the Vardeman and Meeden (1984) stepwise Bayes model.

Next, we define the prior distribution associated with a generic restricted parameter set Ω_{cab}. The reader will not be surprised to see that it is very similar to the prior distribution given in Section 3.1. The basic structure is the following:

$$\Pr(y) = \sum_{h \in \mathcal{H}} \Pr(h) \sum_{r \in \mathcal{R}} \prod_{j=1}^{k_h} \Pr(r_{\mathcal{U}_{jh}} | h) \prod_{u=1}^{c} \Pr(y_{\mathcal{U}_{ur}} | r)$$

The only difference between the prior distribution here and the one in Section 3.1 is that the probability function $\Pr(y|r)$ factors into c independent strata-specific probabilities, where $c \in \{2, \ldots, C\}$, instead of necessarily being equal to two. As usual, $\Pr(h)$ is assumed to be known. The recommendation for how to choose $\Pr(h)$ when k_h varies is the same as that given in Section 2.4. Now, we define the conditional probability function for r.

$$\Pr(r_{\mathcal{U}_{jh}} | h) = \frac{\Gamma(c\epsilon_{jh}) \prod_{z=1}^{c} \Gamma(\epsilon_{jh} + r_z \mathcal{U}_{jh})}{\Gamma(\epsilon_{jh})^c \Gamma(c\epsilon_{jh} + N_{jh})}$$

This function is simply an extension of the one given in Section 3.1 so that $r_i \in \{1, 2, \ldots, c\}$ instead of just $\{0, 1\}$. Also, note that the stepwise Bayes model for r here is almost identical to the stepwise Bayes model for y in Section 2.2. In fact, the only difference is that we assume here that the strata labels are always the integers

$1, 2, \ldots, c$. So, if we regard the observed values in either model as simply labels, statistical inference is the same, and the discussions regarding estimation and hyperparameter choice from Chapter 2 are applicable here. We will hereafter assume that the reader is familiar with the content from Chapter 2, and simply refer to the posterior distribution for r as $\Pr(r_{t'}|r_t)$. This allows us to avoid writing out the notation associated with h because it will be implied in $\Pr(r_{t'}|r_t)$.

Next, we define the condiional probability function for y. For $u = 1, 2, \ldots, c$,

$$\Pr(y_{\mathcal{U}_{ur}}|r) = \frac{\prod_{r\in b_u}\Gamma(y_{r\mathcal{U}_{ur}})}{\Gamma(N_{ur})}$$

This function $\Pr(y_{\mathcal{U}_{ur}}|r)$ is similar to its counterpart from Section 3.1, but differs in an interesting way. The reader will notice that, as mentioned earlier, the η hyperparameter no longer appears. The function that is defined can be obtained by integrating out the mixing parameter from a Dirichlet-Multinomial mixture as we did in Section 2.1 (a Beta-Binomial mixture was used there in particular), but setting the Dirichlet hyperparameter to a vector of zeros. The resulting integral is finite only if each possible value from the Multinomial model (here, each element of b_u) appears at least once. In the set-up used for Section 3.1, this was not guaranteed, but now in the stepwise Bayes model, we have defined the restricted parameter sets so that it is guaranteed. If we consider the resulting conditional prior distribution $y|r$, we find that it is identical to the stepwise Bayes model presented by Vardeman and Meeden (1984) mentioned earlier.

Now, we present the posterior distribution on Ω_{cab}.

$$\begin{aligned}\Pr(y_{s'}|y_s, r_t) &= \sum_{r\in\mathcal{R}}\Pr(r_{t'}|r_t)\Pr(y_{s'}|y_s, r)\\ &= \sum_{r\in\mathcal{R}}\Pr(r_{t'}|r_t)\prod_{u=1}^{c}\left(\frac{\Gamma(n_{ur})\prod_{r\in b_u}\Gamma(y_{r\mathcal{U}_{ur}})}{\Gamma(N_{ur})\prod_{r\in b_u}\Gamma(y_{rs_{ur}})}\right)\end{aligned}$$

This posterior distribution is a finite mixture model where each individual model is the posterior distribution given in Vardeman and Meeden (1984) for a completely known stratification r, and the mixture weights are $\Pr(r_{t'}|r_t)$.

We now explore this posterior distribution by presenting the conditional probability that $y_i = z$ for a unit $i \in s'$ and $z \in B$ given that $r_i = u$ for $u \in \{1, 2, \ldots, c\}$.

$$\begin{aligned}\Pr(y_i = z|y_s, r_t, r_i = u) &= \frac{\Pr(y_i = z, y_{s_{ur}}|r_t, r_i = u)}{\Pr(y_{s_{ur}}|r_t)} \\ &= \frac{\Gamma(y_{zs_{ur}} + 1)\Gamma(n_{s_{ur}})}{\Gamma(y_{zs_{ur}})\Gamma(n_{s_{ur}} + 1)} \\ &= \frac{y_{zs_{ur}}}{n_{s_{ur}}}\end{aligned}$$

Hence, the marginal posterior distribution for a unit $i \in s'$, when $r_i = u$ is either observed or conditioned on, is just the empirical distribution implied by observing $y_{s_{ur}}$, for $u =\in \{1, 2, \ldots, c\}$. Now, we can give the posterior expectation of a unit $i \in s' \cap t$. Suppose that $r_i = u$ for $u \in \{1, 2, \ldots, c\}$, and then

$$\mathrm{E}[y_i|y_s, r_t] = \bar{y}_{s_{ur}}$$

This example shows how, for units in $s' \cap t$, estimation does not utilize $\mathcal{H}$ at all. This is an expected outcome of assuming that $y \perp h\,|r$. When $i \in s' \cap t'$, however, we once again make use of $\mathcal{H}$ because it appears in $\Pr(r_{t'}|r_t)$.

$$\mathrm{E}[y_i|y_s, r_t] = \sum_{r \in \mathcal{R}} \Pr(r_{t'}|r_t)\bar{y}_{s_{r_i r}}$$

We now briefly discuss the stepwise Bayes estimators under squared-error loss (i.e. the posterior expectation) for the population parameter $\gamma(y)$ based on this model and observing the sample (r_t, y_s). First, we write

$$\mathrm{E}[\gamma(y)|(r_t, y_s)] = \sum_{r \in \mathcal{R}} \Pr(r_{t'}|r_t)\mathrm{E}[\gamma(y)|y_s, r]$$

And in the special case where $\gamma(y) = \mu$, we have the estimator

$$\mathrm{E}[\mu|(r_t, y_s)] = \sum_{r \in \mathcal{R}} \Pr(r_{t'}|r_t) \sum_{u=1}^{c} \frac{N_{ur}}{N} \bar{y}_{s_{ur}}$$

We can see that this estimator is just a weighted average of the design-based stratified estimator using different stratifications.

This posterior can also be used to create sampling weights, as in Section 2.6. For $i \in s$, the sampling weight w_i will simply be a weighted average of the design-based sampling weights associated with stratification using each possible $r \in \mathcal{R}$.

$$w_i = \mathrm{E}\left[\sum_{i'=1}^{N} \mathbb{I}(y_{i'} = y_i)|y_s\right] = \sum_{r\in\mathcal{R}} \Pr(r|y_s, r_t)\frac{N_{ur}}{n_{ur}}$$

We provide two more notes on this model. First, regarding admissibility, the restricted sets Ω_{cab} form a partition of the full parameter set Ω, and the prior distribution for each one assigns positive probability to every element in it. Hence, each pair $(r, y) \in \Omega$ has positive probability under exactly one prior distribution, and each possible sample (r_t, y_s) has positive probability under at least one prior distribution. This implies that our model will produce unique stepwise Bayes (admissible) rules under squared-error loss. Second, regarding how a prior distribution will be selected by a sample, the particular ordering of the prior distributions we have used implies that, upon observing y_s, r_t, the "simplest" restricted parameter set consistent with the sample will be selected. By the "simplest" consistent set, we mean that r will be able to take on only the distinct values observed in r_t and that y will be able to take on only the distinct values observed in y_s. But furthermore, whenever $r_i = u$ is given for $i \in s'$, either by observation or conditioning, the possible values that y_i can take will just be those observed in $y_{s_{ur}}$.

3.3 Simulated non-response examples

In this section, we present simulated examples of a multiple stratification approach to a non-response problem on two artificial populations of size $N = 1,000$. For both populations, the basic idea is that we select a sample t of size $m = 100$ using simple random sampling, observe y for a subset of units $s_0 \subset t$, select a sub-sample s_1 of $n_1 = 15$ non-responders, and successfully observe y_i for each unit i in the subsample (unless $t - s_0$ contains less than 15 units, in which case $s_1 = t - s_0$). We then observe y for all units in $s = s_0 \cap s_1$, and r (the "response behavior" of units) for all units in t. In addition, we completely know a set of stratifications $\mathcal{H}$ where we believe that some $h \in \mathcal{H}$ is a good partition of the population in terms of r. For both populations, we use the same three stratifications (we'll recycle the labels h^1, h^2, and h^3 from Section 2.7 here) and the same r, but we use a continuous distribution for y in Population 1 and a discrete distribution for y in Population 2. First, we present the distribution used to

generate r. For $i \in \mathcal{U}$,

$$r_i|h^1 \overset{\text{iid}}{\sim} \begin{cases} \text{Bernoulli}(0), & h_i^1 = 1 \\ \text{Bernoulli}(3/4), & h_i^1 = 2 \end{cases}$$

where $r_i = 0$ means that y_i will respond if selected in the first round of sampling, and $r_i = 1$ means that y_i will not.

Next, we present the distribution used to generate y for the two populations. For $i \in \mathcal{U}$,

Population 1:

$$y_i|r_i = u \overset{\text{iid}}{\sim} \text{Normal}(\theta_u, 1), \theta_0 = 3, \theta_1 = 6$$

Population 2:

$$\begin{aligned} y &\in \{1, 2, \ldots, 10\}^N \\ y_i|r_i = u &\overset{\text{iid}}{\sim} \text{Multinomial}(1, \theta_u) \\ \theta_0 &= (.1, .1, .1, .1, .2, .3, .4, .5, .6, .7) \\ \theta_1 &= (.7, .6, .5, .4, .3, .2, .1, .1, .1, .1) \end{aligned}$$

See also Figures 3.1 and 3.2 for graphical depictions of the populations and their relationship with r, and Section 3.3 for scripts that will generate these populations and stratifications in R Core Team (2013).

The three stratifications were constructed as follows.

$$h_i^1 = \begin{cases} 1, & i = 1, 2, \ldots, 749, 750 \\ 2, & i = 751, 752, \ldots, 999, 1000 \end{cases}$$

$$h_i^2 = \begin{cases} 1, & i = 1, 2, \ldots, 649, 650, 951, 952, \ldots, 999, 1000 \\ 2, & i = 651, 652, \ldots, 949, 950 \end{cases}$$

$$h_i^3 = \begin{cases} 1, & i = 1, 3, \ldots, 997, 999 \\ 2, & i = 2, 4, \ldots, 998, 1000 \end{cases}$$

For each population, we drew 10,000 samples in the manner described above. For each sample, five methods of estimation of μ were calculated. The label SRS represents the simple random sampling estimator, $\bar{y}_s$ (which is a biased estimator that does not

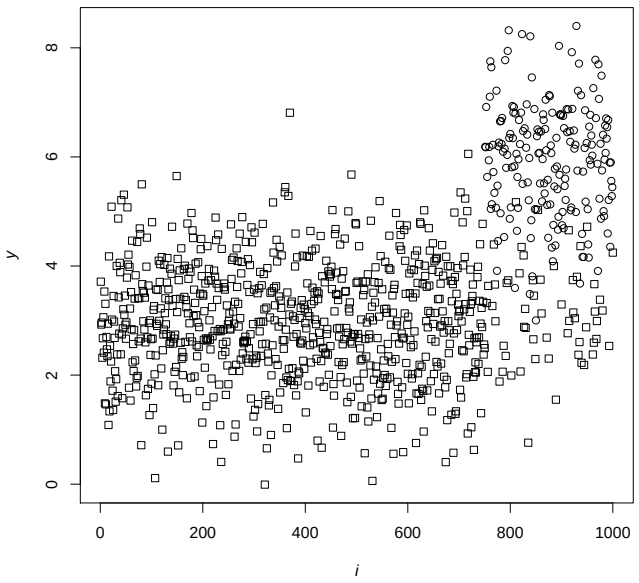

Figure 3.1: Plot of Population 1's y_i for $i \in \mathcal{U}$. The plot characters are a one-to-one function of r.

account for the non-response in any way). The same interval estimator as in Section 2.7 accompanied SRS.

The label $NRSRS$ represents the point estimator

$$\hat{\mu}_{NR} = \bar{y}_{s_{1r}}\bar{r}_t + \bar{y}_{s_{0r}}(1 - \bar{r}_t)$$

which accounts for non-response but does not make use of any information contained in $\mathcal{H}$. This is actually the main estimator proposed in Hansen and Hurwitz (1946). The interval estimator for $NRSRS$ was based on a bootstrap approach. Specifically, for each sample and corresponding estimate $\hat{\mu}_{NR}$, we generated $B = 5,000$ bootstrapped versions of $\hat{\mu}_{NR}$, and took the 0.025 and 0.975 quantiles from this bootstrap distribution as the lower and upper limits, respectively. Each bootstrap version of $\hat{\mu}$ was generated by drawing a standard with-replacement "Efron" bootstrap sample (Efron and Tibshirani, 1994) separately for r_t, $y_{s_{1r}}$ and $y_{s_{0r}}$, and calculating a bootstrap version of $\hat{\mu}_{NR}$ based

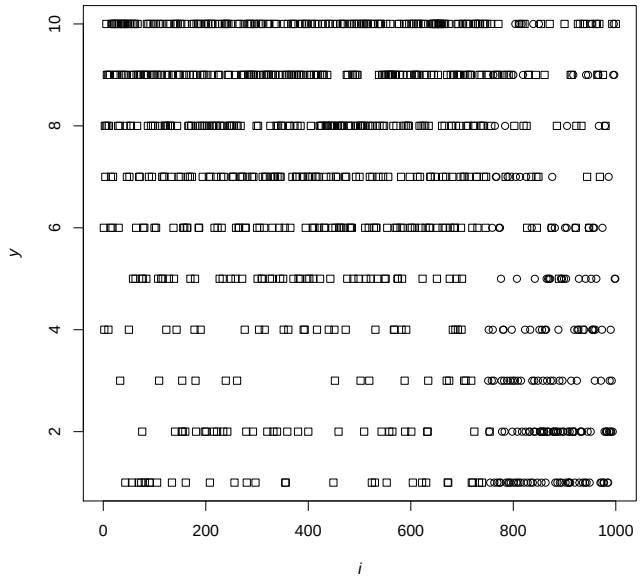

Figure 3.2: Plot of Population 2's y_i for $i \in \mathcal{U}$. The plot characters are a one-to-one function of r. Some "jittering" was used on the vertical axis to make the plot characters more legible.

on them.

The label $NRPS_a$ for $a = 1, 2, 3$ represents the point estimator

$$\hat{\mu}_{NR_a} = \bar{y}_{s_{1r}} \left(\sum_{j=1}^{2} \frac{N_{jh^a}}{N} \bar{r}_{t_{jh^a}} \right) + \bar{y}_{s_{0r}} \left(1 - \left(\sum_{j=1}^{2} \frac{N_{jh^a}}{N} \bar{r}_{t_{jh^a}} \right) \right)$$

The difference between this point estimator and the one for $NRSRS$ is that $\bar{r}_t$ is replaced with a post-stratified estimator of the population mean of r using stratification a. The interval estimators here are also based on a bootstrap approach. Specifically, we generated $B = 5,000$ bootstrapped versions of $\hat{\mu}_{NR_a}$ by drawing with-replacement bootstrap samples separately for $r_{t_{1h^a}}$, $r_{t_{2h^a}}$, $y_{s_{1r}}$ and $y_{s_{0r}}$.

The label MS represents estimation using the multiple stratification model presented in Section 3.2. The hyperparameters for this model were chosen as recommended,

specifically $\epsilon_{jh} \propto N_{jh}$ and $\sum_{j=1}^{k_h} \epsilon_{jh} = 1/10$. The prior distribution $\Pr(h)$ was set to be constant across $\mathcal{H}$. Calculations were done using a Monte Carlo approximation with 1,000 independently drawn samples for the point estimator ($\mathrm{E}[\mu|r_t, y_s]$) and interval estimator (0.025 and 0.975 quantiles of $\mu|r_t, y_s$).

Table 3.1: Simulation results: *mae* gives the mean absolute error of the point estimator, *bias* gives the bias of the point estimator, *lower* gives the mean of the lower interval estimator limit, *width* gives the mean width of the interval estimator, and *cover* gives the coverage probability of the interval estimator.

Population 1	mae	bias	lower	width	cover
$NRPS_1$	0.094	-0.001	3.299	0.484	0.957
MS	0.095	0.001	3.314	0.454	0.942
$NRPS_2$	0.112	-0.001	3.260	0.572	0.952
$NRSRS$	0.122	-0.002	3.238	0.620	0.953
$NRPS_3$	0.122	-0.002	3.238	0.619	0.952
SRS	0.126	-0.109	3.144	0.580	0.944
Population 2	mae	bias	lower	width	cover
$NRPS_1$	0.207	0.002	6.213	1.045	0.954
MS	0.207	0.000	6.240	0.979	0.940
$NRPS_2$	0.218	0.001	6.173	1.118	0.955
$NRSRS$	0.224	0.003	6.151	1.161	0.957
$NRPS_3$	0.225	0.002	6.151	1.161	0.956
SRS	0.228	0.140	6.329	1.100	0.942

The simulation results are given in Table 3.3. For both populations, the relative performance of the design-based estimators (in terms of mean absolute error) follows a similar pattern and is not surprising. $NRPS_1$, which uses h^1 to post-stratify r, is the "correct" estimator, and does the best. $NRPS_2$ and $NRPS_3$ are essentially using a moderately and highly distorted version of h^1 to post-stratify r, respectively, do not perform as well. The performance of $NRPS_3$ suggests that h^3 groups r poorly enough that it offers no performance improvement over $NRSRS$, which ignores all stratification information. Of course, ignoring the non-response, as does SRS, results in the worst performance, and is only case where a substantial bias is observed. Finally, the coverage probabilities of all the design-based interval estimators is approximately equal to the nominal 95% probability. The interval estimators that used the bootstrap technique

may be slightly over-covering because no finite-population correction was used. The mean width of the interval estimators mostly follows the same pattern as the mean absolute error, except that the SRS interval estimator, which was based on Normal theory and not bootstrapping, is not very wide on average.

For both populations, the MS estimator performs (in terms of mean absolute error) approximately as well as $NRPS_1$. In other words, even if we do not know the correct stratification of r, estimation does not suffer. Importantly, the performance of the MS estimator here did not depend on the type of distribution followed by y, which contrasts with the simulated examples of the standard MS estimator in Section 2.7. This is because the stratification mixture weights from the standard MS estimator presented in Chapter 2, $\Pr(h|y_s)$, needed within-stratum homogeneity (in the form of repeated values) in y_s to be present to discern between a good and bad stratification. In contrast, the mixture weights for the MS estimator presented in this chapter, $\Pr(h|r_t)$, only look at r_t, not y_s. Because r_t is the observed response behavior of the units in t, it is always highly discrete, and provides the model with a strong signal about which stratifications group r well. The coverage probability of the MS interval estimator is also approximately equal to the nominal 95% probability for both populations. Its mean width is also the estimator is also the smallest of the five interval estimators for both populations, although this may be partly because the bootstrap technique used for most of the design-based estimators did not use a finite-population correction.

Artificial population generation

Here we provide scripts that can be used to generate the artificial populations and stratifications discussed above in R Core Team (2013).

Population 1:

```
set.seed(1)


## y
theta <- c(rep(1,750),rep(2,250))
f.gen.t <- function(u) ifelse(u==1,0,rbinom(n=1,size=1,prob=.75))
t.pop <- sapply(theta, f.gen.t)
```

```
f.t <- function(t){
  if(t==0) out <- rnorm(1,3)
  if(t==1) out <- rnorm(1,6)
  return(out)
}

y.pop <- sapply(t.pop,f.t);tapply(y.pop,theta,mean)
N <- length(y.pop)

## stratifications
h.perfect <- theta
h.good <- c(rep(1,650),rep(2,300),rep(1,50))
h.bad <- rep(1:2,500)
H <- data.frame(h.perfect,h.good,h.bad)
```

Population 2:

```
set.seed(1)

## y
theta <- c(rep(1,750),rep(2,250))
f.gen.t <- function(u) ifelse(u==1,0,rbinom(n=1,size=1,prob=.75))
t.pop <- sapply(theta, f.gen.t)

f.t <- function(t){
  if(t==0) out <- sample(1:10,1,prob=c(1,1,1,1:7))
  if(t==1) out <- sample(1:10,1,prob=c(7:1,1,1,1))
  return(out)
}

y.pop <- sapply(t.pop,f.t);tapply(y.pop,theta,mean)
N <- length(y.pop)
```

```
## stratifications
h.perfect <- theta
h.good <- c(rep(1,650),rep(2,300),rep(1,50))
h.bad <- rep(1:2,500)
H <- data.frame(h.perfect,h.good,h.bad)
```

3.4 Further developments in handling non-response

Randomly subsampling non-respondents using some sort of followup survey is perhaps the "gold standard" for dealing with non-response. It ensures that our sample of the response is representative of the entire population, and avoids having to make any major assumptions about the relationship between r and y. However, in many practical applications, it may unrealistic to achieve zero non-response in the final round of sampling. As mentioned at the beginning of this chapter, such a "hard core" of non-responders necessitates using a model for non-response. In the simplest case, this model may just be that, for the segment of the population sampled in the final round, non-responders are assumed to be exchangeable with responders. Commonly, though, some sort of auxiliary information is used to model non-response.

Little and Rubin (2002) provide a good description of this non-response problem. In their set-up, the mechanism that creates non-response (or any missing data) falls into one of three cases: MCAR, where data are "missing completely at random"; MAR, where data are "missing at random"; and NI, where the missingness mechanism is "nonignorable". These assumptions can be defined from a frequentist perspective, but also work nicely from a Bayesian perspective, so that is how we use them.

MCAR is the restrictive assumption, and means that y and r are independent. If we assume that a MCAR mechanism is generating non-response, we could even go so far as to act as if the respondent sample units were the entire sample, and ignore the fact that some units didn't respond. If two- (or more) phase sampling is used to decrease non-response, a MCAR assumption could also be employed as in the example given in the previous paragraph where the "hard core" non-respondents are assumed to be exchangeable with those units that do respond in the final round of sampling. NI is

the least restrictive assumption and supposes that, even after conditioning on available information, there may be some residual relationship between y and r. In this case, the statistician essentially must know the relationship between y and r from some source independent of the survey data or simply accept an unknown amount of bias.

Practical applications typically make use of the MAR assumption, which represents a moderate choice between the extremes of MCAR and NI. MAR makes use of some auxiliary information, denoted x, and assumes that $y \perp r \,|x$. Suppose that we conduct only one round of sampling and have some non-response. If we assume that, for some stratification h^*, the MAR assumption holds (i.e. $y \perp r \,|h^*$), we could post-stratify on h^* and ignore the non-response. Alternatively, suppose that we believe h^* is contained in the set of stratifications $\mathcal{H}$, but we do not know which one it is. If we make the additional assumption that h^* does the best job of segmenting y out of all stratifications in $\mathcal{H}$, and y is fairly discrete, we could use the stepwise Bayes model from Chapter 2, and reasonably expect that bias due to non-response would be minimized. An interesting extension to this line of thinking to pair a similar MAR assumption with the use of follow-up survey data. Specifically, suppose that we have the same set-up for data as in this chapter, but that there is still some non-response after the follow-up. Let $r_i = 2$ for a unit i that does or would refuse to respond to both rounds of sampling. If we ignore the non-response in the second round of sampling and use the stepwise Bayes presented in this chapter, we are making an MCAR type of assumption that the one-time non-responders and the two-time non-responders are exchangeable. On the other hand, we could consider a conditional distribution for the units in $y_{\mathcal{U}_{1r}} \cup y_{\mathcal{U}_{2r}}$ (i.e. all one- and two-time non-responders in the population) that depends on some additional auxiliary information. For example, there may be some known h^* where we assume that the distribution of $y_i|h^*$ is the same for units i where $r > 0$, and then, within this set of units, estimation can be performed independently for each strata defined by h^*. Alternatively, like the example above, we may believe that h^* lies in some set $\mathcal{H}_2$ (which may or may not contain the same stratifications as $\mathcal{H}$), and a multiple stratification approach could be developed. In this strategy, $\mathcal{H}$ would be used to improve estimation of r and $\mathcal{H}$ would be used to improve estimation of $y_{\mathcal{U}_{1r}} \cup y_{\mathcal{U}_{2r}}$ or possibly all of y. Future research on applying the multiple stratification approach to non-response problems could focus on developing an effective stepwise Bayes model where auxiliary information is used to

improve estimation of both r and y.

Chapter 4

Software

In this chapter, we describe some functions that can be used to implement our methods in the R statistical computing language (R Core Team, 2013). First, we provide user guide to the functions, and then the actual function definitions are given.

4.1 User Guide

4.1.1 `ds.mult.strat`

This function can be used to conduct estimation based on the stepwise Bayes model presented in Section 3.2. It calculates the posterior distribution $\Pr(h|x_t)$ exactly, and draws a sample from the posterior distribution $\texttt{fun}(y)|y_s, x_t$ where `fun` is a user-defined function of the complete population y.

Arguments

- `H`: A `data.frame` with N rows where each column is a stratification of the population.
- `t`: The partially-observed stratification vector for the entire population. Unobserved units should take the value `NA`.
- `y`: The response vector for the entire population. Unobserved units should take the value `NA`.

- `R`: The number (nonnegative integer) of realizations to sample from the posterior of `fun`. The default value is zero.
- `R.prior`: The number (nonnegative integer) of realizations to use in the Monte Carlo approximation for `adjust.prior`. Defaults to 10,000.
- `fun`: A function of y whose posterior distribution will be sampled. Defaults to the population mean.
- `eps.scale`: A positive real-number equal to the sum of the hyperparameters $\epsilon_1, \epsilon_2, \ldots, \epsilon_{k_h}$ for each $h \in \mathcal{H}$.
- `h.prior`: A vector of length `ncol(H)` defining the (unnormalized) user-supplied prior distribution $\Pr(h)$ for $h \in \mathcal{H}$.
- `adjust.prior`: Logical scalar. See Details.

Details

Each column of H represents a stratification, h. For each h, the hyperparameters $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{k_h})$ are set to sum to `eps.scale`, and to be proportional to $(N_1, N_2, \ldots, N_{k_h})$, i.e. the population stratum sizes with respect to the given h. The posterior distribution of `fun`$(y)|y_s$ is sampled in three steps. First, h is sampled from the distribution $\Pr(h|y_s)$. Then, x is sampled from the distribution $x|x_t, h$. The function `mult.strat` is used to complete these two steps. Finally, y is sampled from the distribution $y|y_s, x$ (using the function `stpolyap` and its image under `fun` is calculated.

If the stratifications in `H` have varying numbers of strata in them, `adjust.prior` should be used. This will cause the prior supplied in `h.prior` to be adjusted as described for discrete responses (since `t` will always be discrete) in Section 2.4.

Value

This function returns a list with either one or two components. A vector `h.post` of length `ncol(H)` which contains the posterior distribution $\Pr(h|y_s)$ is always returned. If a positive value is supplied for `R`, then a vector `fun.sim`, containing the sample of size `R` from the posterior distribution `fun`$(y)|y_s, x_t$ is also returned.

4.1.2 `mb.sample`

This function implements the multi-balanced sampling approach described in Section 2.5.

Arguments

The user supplies the following arguments:

- `H`: A matrix of data.frame where each column gives the stratum membership for some stratification
- `n`: the size of the sample to be drawn
- `m`: the (maximum) number of samples to be drawn (when the `'at.least'` method is used)
- `use.p` a logical scalar indicating whether or not to incorporate a user-supplied prior distribution in the loss function
- `p.h`: a vector the same length as `ncol(H)` giving the (unnormalized) prior distribution on `H`
- `method`: one of `'find.best'`, `'how.many'`, or `'at.least'`; see Details section
- `max.loss`: for the `'how.many'` method, the maximum loss function value for a sample to be considered a possibility; for the `'at.least'` method, the maximum loss function for a sample to be returned as successful.

Details

This function implements the multi-balanced sampling mechanism described in Section 2.5. For each of three methods used in this function, simple random samples of size `n` are repeatedly drawn from the population using the R `sample` function. On each draw, requirement **mb2** is checked (**mb1** is always fulfilled). If it is fulfilled, the sample's incurred loss as described in **mb3** is checked.

The `'find.best'` method attempts to fulfill **mb3** by returning the sample with smallest incurred loss out of `m` draws generated as described above. The `'how.many'`

method informs the user about the number of samples that fulfill **mb1**, **mb2**, and **mb3'** for a user-supplied `max.loss` by drawing `m` samples as described above, and recording how many unique samples fulfilled **mb3'** and estimating the total number of samples that fulfill these three requirements by multiplying the ratio of successful draws (not just the number of unique successful draws) over `m` by the total number of possible draws, i.e. $\binom{N}{n}$. The `'at.least'` attempts to return a sample that fulfills **mb1**, **mb2**, and **mb3'** for a user-supplied `max.loss`. This method generates samples as described above until it finds one, or until `m` samples have been generated.

Value

For the `'find.best'` method, the following objects are returned:

- `s`: The sample that minimized the loss function from **mb3** out of `m` attempts.
- `loss`: The incurred loss of `s`.
- `alloc`: A list with `ncol(H)` components. The j^{th} component is a `data.frame` with as many rows as the number of strata defined by the j^{th} column of `H`. The first column gives the number of sample units in each stratum, and the second column gives the proportion of population units that lie in each strata.

For the `'how.many'` method, the following objects are returned:

- `at.least`: The number of unique samples that fulfilled **mb1**, **mb2**, and **mb3'** in `m` attempts.
- `total.est`: An estimate of the number of samples that fulfill **mb1**, **mb2**, and **mb3'**.
- `possible.alloc`: The same list as `alloc` except corresponding to just one of the `at.least` samples found.

If the `'at.least'` method if successful, it is returns the same objects as the `'find.best'`. If not, it returns the character string `'failed to find an acceptable sample'`).

4.1.3 `mult.strat`

This function can be used to conduct estimation based on the stepwise Bayes model presented in Section 2.2. It calculates the posterior distribution $\Pr(h|y_s)$ exactly, and draws a sample from the posterior distribution $\texttt{fun}(y)|y_s$ where `fun` is a user-defined function of the complete population y.

Arguments

- `H`: A `data.frame` with N rows where each column is a stratification of the population.
- `y`: The response vector for the entire population. Unobserved units should take the value `NA`.
- `R`: The number (nonnegative integer) of realizations to sample from the posterior of `fun`. The default value is zero.
- `R.prior`: The number (nonnegative integer) of realizations to use in the Monte Carlo approximation for either `adjust.prior.discrete` or `adjust.prior.continuous`. Defaults to 10,000.
- `fun`: A function of y whose posterior distribution will be sampled. Defaults to the population mean.
- `eps.scale`: A positive real-number equal to the sum of the hyperparameters $\epsilon_1, \epsilon_2, \ldots, \epsilon_{k_h}$ for each $h \in \mathcal{H}$.
- `h.prior`: A vector of length `ncol(H)` defining the (unnormalized) user-supplied prior distribution $\Pr(h)$ for $h \in \mathcal{H}$.
- `adjust.prior.discrete`: Logical scalar. See Details.
- `adjust.prior.continuous`: Logical scalar. See Details.

Details

Each column of H represents a stratification, h. For each h, the hyperparameters $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{k_h})$ are set to sum to `eps.scale`, and to be proportional to $(N_1, N_2, \ldots, N_{k_h})$,

i.e. the population stratum sizes with respect to the given h. The posterior distribution $h|y_s$ is first calculated, partly with the help of the function `ys.given.h`, which calculates $\Pr(y_s|h)$. Then, the posterior distribution of `fun`$(y)|y_s$ is sampled in two steps: h is sampled from the distribution $\Pr(h|y_s)$, and then y is sampled from the distribution $y|y_s, h$ and its image under `fun` is calculated.

If the stratifications in `H` have varying numbers of strata in them, either `adjust.prior.discrete` or `adjust.prior.continuous` should be used depending on the type of response that `y` is. Use the continuous version if every value observed in `y` is distinct, and the discrete version otherwise. This will cause the prior supplied in `h.prior` to be adjusted as described in Section 2.4.

Value

This function returns a list with either one or two components. A vector `h.post` of length `ncol(H)` which contains the posterior distribution $\Pr(h|y_s)$ is always returned. If a positive value is supplied for `R`, then a vector `fun.sim`, containing the sample of size `R` from the posterior distribution `fun`$(y)|y_s$ is also returned.

4.1.4 `stpolyap`

This function takes a (sample of a) stratified population, and simulates a complete copy of it using the stratified Polya posterior (Vardeman and Meeden, 1984).

Arguments

- `y`: The response vector for the entire population. Unobserved units should take the value `NA`.
- `h`: A stratification of the population. It has the same length as `y`, and provides the stratum membership of each unit.
- `fun`: A function that takes a vector of length `N` (a complete copy of `y`) as its input. Its default value is the identity function.

Details

`stpolyap` uses the function `polyap` separately on each stratum to generate realizations of the unobserved units.

Value

The image under `fun` of the simulated complete copy of y is returned.

4.2 Function definitions

4.2.1 `ds.mult.strat`

```
ds.mult.strat <- function(H, t, y,R,R.prior=1e4,
                          fun=function(u)mean(u),
                          eps.scale=1/10,
                          h.prior=rep(1,ncol(H)),
                          adjust.prior=FALSE){
  ## set up objects and define variables
  L <- ncol(H) # number of stratifications being considered
  H.s1 <- data.frame(H[!is.na(t),]) # rows of H in s1
  t.s1 <- t[!is.na(t)] # t with NAs omitted
  b <- sort(unique(t.s1)) # restricted parameter space for t
  lp.ts1.h <- rep(0,L) # placeholder
  post.eps <- list()
  n1 <- length(t.s1)
  N <- length(t)

  ## create eps matrices to favor prop alloc if only eps.scale
  eps <- lapply(H,function(h){
    stratum.weights <- table(h)/length(h)
    eps.h <- sapply(stratum.weights,
                    function(w) rep(w*eps.scale,length(b)))
    rownames(eps.h) <- b
```

```
  return(eps.h)
})

## adjust prior for varying k (number of strata)
## only discrete version will be applicable
lp.h <- log(h.prior)
if(adjust.prior.discrete){
  a <- length(b)
  k <- sapply(H,function(h) length(unique(h)))
  H.base <- sapply(k,function(k.h) rep(1:k.h,length.out=N))
  eps.base <- lapply(data.frame(H.base),function(h){
    stratum.weights <- table(h)/length(h)
    eps.h <- sapply(stratum.weights,
                    function(w) rep(w*eps.scale,length(b)))
    rownames(eps.h) <- b
    return(eps.h)
  })
  p.base.sample <- t(sapply(1:R.prior,function(u){
    t.base <- sample(b,N,replace=T)
    s <- sample(N,n1)
    lp.base <- ys.given.h(H.base[s,], t.base[s], eps.base, b,
                          ret.post.eps=FALSE)
    lp.base.adj <- lp.base-max(lp.base)
    return(exp(lp.base.adj)/sum(exp(lp.base.adj)))
  }))
  lp.k <- -log(apply(p.base.sample,2,mean))
  lp.h <- lp.h+lp.k;lp.h
}

## calculate log probability of t.s1 given h for each h
post.calc <- ys.given.h(H.s1, t.s1, eps, b)
lp.ts1.h <- post.calc[[1]]
```

```
post.eps <- post.calc[[2]]

## calculate posterior probabilities
lupost <- lp.h + lp.ts1.h
lupost.adj <- lupost - max(lupost) # this avoids sum(exp(lupost))=0
                                   # in large samples
h.post <- exp(lupost.adj)/sum(exp(lupost.adj));h.post

## draw from posterior
which.h <- sample(1:ncol(H),size=R,prob=h.post,replace=T)
fun.sim <- NULL

for(h.ind in which.h){
  ## draw a probability vector from the posterior
  gamma.sim <- apply(post.eps[[h.ind]],2, function(e.vec)
                     sapply(e.vec, function(e) rgamma(1,e)))
  dir.sim <- apply(gamma.sim, 2, function(g) g/sum(g))

  ## generate a full t conditional on probability vector
  t.copy <- t
  t.copy[is.na(t)][sort.list(H[,h.ind][is.na(t)])] <-
    unlist(sapply(unique(H[,h.ind]), function(j)
                  sample(b,sum(is.na(t[H[,h.ind]==j])),
                         replace=TRUE,
                         dir.sim[,colnames(dir.sim)==j])))

  ## generate a full y conditional on t
  if(!is.null(fun.sim)){
    fun.sim <- data.frame(fun.sim,stpolyap(y,t.copy,fun))
  } else {
    fun.sim <- stpolyap(y,t.copy,fun)
  }
```

```
  }

  ## return posterior on H, and sample of full y's
  names(fun.sim) <- which.h
  out <- list(h.post,fun.sim)
  names(out) <- c('h.post','fun.sim')
  return(out)
}
```

4.2.2 mb.sample

```
mb.sample <- function(H, # data.frame of stratifications
                      n, # sample size
                      m, # max number of tries for at.least method
                      use.p=FALSE, # should p.h be used to define loss?
                      p.h=NULL, # prior on H
                      method='find.best', # find.best returns the best
                                          # sample observed in m tries,
                                          # how.many returns the number
                                          # of samples with a loss <=
                                          # max.loss seen in m tries,
                                          # and at.least returns a
                                          # sample with loss <= max.loss
                      max.loss=NULL){
    ## figure out allocation target
    N <- nrow(H)
    W.h <- apply(H,2,function(u) tapply(u,u,function(v) length(v)/N))
                                          # pop str weight for each h

    desired.n.h <- lapply(W.h,function(u) u*n)
                                          # sample allocation for each h
                                          # under proportionally-alloc
                                          # stratified random sampling
```

```
## format weights from p.h (or just a bunch of 1's if use.p=FALSE)
if(use.p){
    length.h <- sapply(desired.n.h,length)
    weights.h <- c()
    for(i in 1:length(H))
        weights.h <- c(weights.h,rep(p.h[i],length.h[i]))
} else {weights.h <- 1}

## define function that returns allocation wrt each h in H given s
f.n.h <- function(s){
    lapply(H,function(h.arg){tapply(1:N,h.arg,function(u)
                                    length(intersect(u,s)))
                        })
}

if(method=='find.best'){
    ## initialize find.best objects
    best.s <- NULL # this will be returned as the best sample
    best.ss <- NULL # this will be used to keep track of the loss
                    # incurred by the currently best.s

    ## search for most balanced sample
    for(i in 1:m){
        s <- sort(sample(N,n)) # draw a sample
        n.h <- f.n.h(s) # calculate allocations
        check.2min <- min(unlist(n.h))>1
            # does each stratum have >= 2?
        if(check.2min==FALSE) {ss <- Inf} else {
            ss <- sum(weights.h*(unlist(desired.n.h)-unlist(n.h))^2)
                # evaluate loss
        }
```

```
        if(is.null(best.s)) {best.s <- s;best.ss <- ss} else {
            if(ss<best.ss) {best.s <- s;best.ss <- ss}
        }
    }

    ## set up output for find.best method
    alloc <- list()
    achieved.n.h <- f.n.h(best.s)
    for(i in 1:length(desired.n.h)) {
        alloc.i <- data.frame(achieved.n.h[[i]],desired.n.h[[i]])
        names(alloc.i) <- c('achieved','desired')
        alloc[[i]] <- alloc.i
    }
    names(alloc) <- names(H)

    out <- list(best.s,best.ss,alloc)
    names(out) <- c('s','loss','alloc')
    return(out)
}
if(method=='how.many'){
    ## initialize how.many objects
    possible.s <- list()
    successes <- 0
    ind <- 1

    ## search for possible samples
    for(i in 1:m){
        s <- sort(sample(N,n)) # draw a sample
        n.h <- f.n.h(s) # calculate allocations
        check.2min <- min(unlist(n.h))>1 # each stratum have >= 2?
        is.possible <-
            (check.2min==TRUE &
```

```
            max.loss>=sum(weights.h*(unlist(desired.n.h)-
            unlist(n.h))^2))  # determine is s is a possible one
        successes <- successes+is.possible
        if(is.possible &
           !any(sapply(possible.s,function(u) all(s==u)))){
              # check to see if this one was already used
            possible.s[[ind]] <- s
            ind <- ind+1
        }
    }

    ## calculate total estimate
    l.prop.est <- log(successes) - log(m)
    total.est <- exp(l.prop.est + lchoose(N,n))

    ## set up output for how.many method
    achieved.n.h <- f.n.h(possible.s[[1]])
    possible.alloc <- list()
    for(i in 1:length(desired.n.h)) {
        alloc.i <- data.frame(achieved.n.h[[i]],desired.n.h[[i]])
        names(alloc.i) <- c('achieved','desired')
        possible.alloc[[i]] <- alloc.i
    }
    names(possible.alloc) <- names(H)

    out <- list(at.least,total.est,possible.alloc)
    names(out) <- c('at.least','total.est','possible.alloc')
    return(out)
}
if(method=='at.least'){
    ## search for acceptable sample
    keeper <- NULL
```

```
    for(i in 1:m){
        s <- sort(sample(N,n)) # draw a sample
        n.h <- f.n.h(s) # calculate allocations
        check.2min <- min(unlist(n.h))>1 # each stratum have >= 2?
        if(check.2min==FALSE) {ss <- Inf} else {
            ss <- sum(weights.h*(unlist(desired.n.h)-unlist(n.h))^2)
                # evaluate loss
        }
        if(ss<=max.loss){
            keeper <- s
            keeper.ss <- ss
            break
        }
    }

    ## set up output for at.least method
    if(!is.null(keeper)){
        alloc <- list()
        achieved.n.h <- f.n.h(keeper)
        for(i in 1:length(desired.n.h)) {
            alloc.i <- data.frame(achieved.n.h[[i]],
                                  desired.n.h[[i]])
            names(alloc.i) <- c('achieved','desired')
            alloc[[i]] <- alloc.i
        }
        names(alloc) <- names(H)

        out <- list(keeper,keeper.ss,alloc)
        names(out) <- c('s','loss','alloc')
        return(out)
    } else {return('failed to find an acceptable sample')}
}
```

```
}
```

4.2.3 mult.strat

```
mult.strat <- function(H, y, R=0, R.prior=1e4,
                       fun=function(u)mean(u),
                       eps.scale=1/10,
                       h.prior=rep(1,ncol(H)),
                       adjust.prior.discrete=FALSE,
                       adjust.prior.continuous=FALSE){

    ## set up objects and define variables
    L <- ncol(H) # number of stratifications being considered
    H.s <- data.frame(H[!is.na(y),]) # rows of H in s
    y.s <- y[!is.na(y)] # y with NAs ommitted
    b <- sort(unique(y.s)) # restricted parameter space for y
    n <- length(y.s)
    N <- length(y)

    ## create eps matrices to favor prop alloc if only eps.scale
    eps <- lapply(H,function(h){
      stratum.weights <- table(h)/length(h)
      eps.h <- sapply(stratum.weights,
                      function(w) rep(w*eps.scale,length(b)))
      rownames(eps.h) <- b
      return(eps.h)
    })

    ## adjust prior for varying k (number of strata)
    ## discrete version
    lp.h <- log(h.prior)
    if(adjust.prior.discrete){
      a <- length(unique(y))-1
```

```
  k <- sapply(H,function(h) length(unique(h)))
  H.base <- sapply(k,function(k.h) rep(1:k.h,length.out=N))
  eps.base <- lapply(data.frame(H.base),function(h){
    stratum.weights <- table(h)/length(h)
    eps.h <- sapply(stratum.weights,
                    function(w) rep(w*eps.scale,length(b)))
    rownames(eps.h) <- b
    return(eps.h)
  })
  p.base.sample <- t(sapply(1:R.prior,function(u){
    y.base <- sample(b,N,replace=T)
    s <- sample(N,n)
    lp.base <- ys.given.h(H.base[s,], y.base[s], eps.base, b,
                          ret.post.eps=FALSE)
    lp.base.adj <- lp.base-max(lp.base)
    return(exp(lp.base.adj)/sum(exp(lp.base.adj)))
  }))
  lp.k <- -log(apply(p.base.sample,2,mean))
  lp.h <- lp.h+lp.k
}
## continuous version
if(adjust.prior.continuous){
  k <- sapply(H,function(h) length(unique(h)))
  H.base <- sapply(k,function(k.h) rep(1:k.h,length.out=N))
  eps.base <- lapply(data.frame(H.base),function(h){
    stratum.weights <- table(h)/length(h)
    eps.h <- sapply(stratum.weights,
                    function(w) rep(w*eps.scale,length(b)))
    rownames(eps.h) <- b
    return(eps.h)
  })
  p.base.sample <- t(sapply(1:R.prior,function(u){
```

```
    s <- sample(N,n)
    lp.base <- ys.given.h(H.base[s,], b, eps.base, b,
                          ret.post.eps=FALSE)
    lp.base.adj <- lp.base-max(lp.base)
    return(exp(lp.base.adj)/sum(exp(lp.base.adj)))
  }))
  lp.k <- -log(apply(p.base.sample,2,mean))
  lp.h <- lp.h+lp.k
}

## calculate log probability of y.s given h and parms for posterior
post.calc <- ys.given.h(H.s, y.s, eps, b)
lp.ys.h <- post.calc[[1]]
post.eps <- post.calc[[2]]

## calculate posterior probabilities
lupost <- lp.h + lp.ys.h
lupost.adj <- lupost - max(lupost) # this avoids sum(exp(lupost))=0
                                   # in large samples
h.post <- exp(lupost.adj)/sum(exp(lupost.adj));h.post

## draw sample if requested
if(R>0){
    which.h <- sample(1:ncol(H),size=R,prob=h.post,replace=T)
    fun.sim <- NULL

    for(h.ind in which.h){
        ## draw a probability vector from the posterior
        gamma.sim <- apply(post.eps[[h.ind]],2, function(e.vec)
            sapply(e.vec, function(e) rgamma(1,e)))
        dir.sim <- apply(gamma.sim, 2, function(g) g/sum(g))
```

```
            ## generate a full y conditional on probability vector
            y.copy <- y
            y.copy[is.na(y)][sort.list(H[,h.ind][is.na(y)])] <-
                unlist(sapply(unique(H[,h.ind]), function(j)
                              sample(b,sum(is.na(y[H[,h.ind]==j])),
                                     replace=TRUE,
                                     dir.sim[,colnames(dir.sim)==j])))
            fun.sim.add <- fun(y.copy)

            ## add to fun.sim object
            if(!is.null(fun.sim))
                fun.sim <- data.frame(fun.sim,fun.sim.add)
            if(is.null(fun.sim))
                fun.sim <- fun.sim.add
        }

        # return posterior on H, and sample of full y's
        names(fun.sim) <- which.h
        out <- list(h.post,fun.sim)
        names(out) <- c('h.post','fun.sim')
        return(out)
    }

    ## otherwise, just return posterior on H
    if(R==0) return(h.post)
}
```

4.2.4 stpolyap

```
stpolyap <- function(y,h,fun=function(u) u){
  y.smp <- tapply(y, h, function(u) u[!is.na(u)])
                                        # separate y into sampled
                                        # observations from each
```

```
                                        # stratum
  y.n <- tapply(y, h, function(u) sum(!is.na(u)))
                                        # find sample size in each
                                        # stratum
  y.N <- tapply(y, h, length) # find pop size in each stratum
  z <- !is.na(y) # logical vector saying which are in s
  unique.h <- sort(unique(as.matrix(h)))

  y.complete <- y # y.complete's NAs will be filled in one stratum at
                  # a time
  for(j in 1:length(y.smp)){
      y.complete[h==unique.h[j]&!z] <-
          polyap(y.smp[[j]],y.N[j]-y.n[[j]])[-(1:y.n[j])]
  }
  return(fun(y.complete))
}
```

4.2.5 ys.given.h

Note: this function is not intended to be called by the user, but must be defined before defining `ds.mult.strat` or `mult.strat`.

```
ys.given.h <- function(H.s, y.s, eps, b, ret.post.eps=TRUE){

  ## set up output objects
  post.eps <- list()
  lp.ys.h <- rep(NA,ncol(H.s))

  ## calculate log probability of y.s given h for each h
  for(h.ind in 1:ncol(H.s)){
    h.s <- H.s[,h.ind]
    ys.tab <- sapply(sort(unique(h.s)),
                     function(j) sapply(b,function(z)
```

```
                                        sum(y.s[h.s==j]==z)))
    colnames(ys.tab) <- sort(unique(h.s))
    rownames(ys.tab) <- b;ys.tab
    post.eps[[h.ind]] <- eps[[h.ind]]+ys.tab
    lp.ysj.h <- sapply(1:length(unique(h.s)),function(j){
      sum(lgamma(sum(eps[[h.ind]][,j])),
          sum(lgamma(post.eps[[h.ind]][,j])),
          -lgamma(sum(post.eps[[h.ind]][,j])),
          -sum(lgamma(eps[[h.ind]][,j])))})
    lp.ys.h[h.ind] <- sum(lp.ysj.h)
  }

  if(ret.post.eps) return(list(lp.ys.h,post.eps)) else return(lp.ys.h)
}
```

References

Basu, D. (1969). Role of sufficiency and likelihood principles in sample survey theory. *Sankyā B*, 31:441–454.

Breidt, F. (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics*, 36:403–427.

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

Dahlke, M., Breidt, F., Opsomer, J., and I., V. K. (2013). Nonparametric endogenous post-stratification in surveys. *Statistica Sinica*, 23:189–211.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman and Hall.

Ericson, W. (1969a). A note on the posterior mean. *Journal of the Royal Statistical Society, Series B*, 31:332–334.

Ericson, W. (1969b). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31:195–233.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume I. Wiley.

Fraley, C. and Raftery, A. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41:587–588.

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

Gelman, A. (2007). Struggles with survey weights and regression modeling. *Statistical Science*, 22:153–164.

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

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

Hansen, M. and Hurwitz, W. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14:332–336.

Hansen, M. and Hurwitz, W. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41:517–529.

Holt, D. and Smith, T. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142:33–46.

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

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London: Series A*, 186:453–461.

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

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. Wiley, 2nd edition.

Mazloum, R. and Meeden, G. (1987). Using the stepwise bayes technique to choose between experiments. *Annals of Statistics*, 15:269–277.

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. (2011). Impact of frequentist and bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26:240–256.

Royall, R. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71:657–664.

Royall, R. (1988). *Handbook of Statistics 6: Sampling*, chapter The Prediction Approach to Sampling Theory. North-Holland.

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

Ruggles, S., Alexander, J., Genadek, K., Goeken, R., Schroeder, M., and Sobek, M. (2010). *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]*. University of Minnesota, Minneapolis.

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

Slud, E. and Thibaudeau, Y. (2010). Simultaneous calibration and nonresponse adjustment. Technical Report Statistics #2010-03, Statistics Research Division, U.S. Census Bureau.

Strief, J. and Meeden, G. (2013). Objective stepwise bayes weights in survey sampling. *Survey Methodology*, 39:1–27.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63:411–423.

Valliant, R., Dorfman, A., and Royall, R. (2000). *Finite Population Sampling and Inference*. Wiley.

Vardeman, S. and Meeden, G. (1984). Admissible estimators in finite population sampling employing various types of prior information. *Annals of Statistics*, 12:675–684.