BAYESIAN PREDICTIVE INFERENCE WITH SURVEY WEIGHTS

by
Lingli Yang

A Thesis

Submitted to the
Department of Mathematical Sciences
In partial fulfillment of the requirement
For the degree of
Master of Science in Applied Statistics
at
Worcester Polytechnic Institute
by

May, 2021

APPROVED:

Thesis Advisor: Balgobin Nandram, Ph.D.

Acknowledgements

Throughout the writing of my thesis, I have received a great deal of assistance and support.

I would first like to thank my thesis advisor, Dr. Balgobin Nandram, whose expertise on Bayes Statistics was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to thank my master program advisor, Dr. Thelge Buddika Peiris, for your patient support and for all of the opportunities I was given to further my research.

I would also like to acknowledge my friends, Ashley Lockwood, Xinyu Chen, Yang Liu, and Zihang Xu, for their wonderful discussions in our weekly seminar. You provided me with the tools that I needed to choose the right direction and successfully complete my thesis.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. Finally, I could not have completed this thesis without the support of my fiance, Kexuan Li, who provided stimulating discussions as well as joyful distractions to rest my mind outside of my research.

Abstract

Sample surveys play a significant role in obtaining reliable estimators of a finite population. In a world of "big data", a large amount of available non-probability samples are easier and faster to obtain than probability samples. In this research, we focus on binary data which occur in many different situations. The main idea of this research is to compare the performance of nine methods with different constructed survey weights, and we can use these methods for non-probability sampling after weights are estimated (e.g. quasi-randomization). In particular, we employ original weights, adjusted weights, adjusted standardized weights, and trimmed weights to build posterior distributions. We apply our models to the simulation study and compare their performance by posterior mean, posterior standard deviation, relative bias, posterior root mean squared error, and the coverage rate of 95% credible intervals. Also, we discuss an application on body mass index and compare these nine models.

Keywords: Bayesian statistics, Non-probability samples, Normalized likelihood, Selection bias, Survey weights.

Table of Contents

Table of Contents (Continued)

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Sample surveys play a significant role in obtaining reliable estimators of finite population, such as means, totals, and ratios. The sample is a subset of a population and a perfect sample would be represented in the sense that our interests in the population could be estimated from the sample with a known degree of accuracy. In an ideal situation, the sampled population where the sample was taken would be identical to the target population. However, the ideal survey is hard to attain. Moreover, according to the items of interest, a good sample should have accurate responses. But in many surveys, it is challenging to collect accurate responses. For instance, the surveys of people usually obtain the sampled population which would be smaller than the target population, and sometimes, people refuse to tell the truth or they do not always understand the questions (Lohr 2009). To simplify the presentation of concepts, we assume the sampled population is the target population, which could be considered as population.

Let y_i, $i = 1, \ldots, N$, be the variable of interest of the i-th unit, where $N < \infty$ is the total number of units in the population. In probability sampling, the probability of each unit being selected is equal or unequal, like simple random sampling, stratified sampling, cluster sampling, and systematic sampling.

Without any doubt, probability sampling is the golden rule for finite population prediction and inference. However, the sampling probability preference is challenged because of nonresponse, time, and cost. The survey response is declining steadily; rare events, such as crashes, need long-term observation; convenience samples are faster, easier, and cheaper to collect; massive data are increasingly available but unstructured and hard to analyze

(Beaumont 2020; Rao 2020).

Therefore, it is necessary to pay more attention to the non-probability sampling. Compared to probability sampling , the non-probability sampling is more feasible considering the desire for access to "real-time" data, the high cost of data collection, and the decline in the response rates (Groves 2011, Miller 2017, Johnson and Smith 2017, Beaumont 2020), Rao et al. (2010) and Haziza, Beaumont, et al. (2017) summarized the typical weighting process in the multipurpose surveys.

Nandram (2007) discussed the Bayesian prediction inference under informative sampling via surrogate samples. In Chapter 2, we introduce the surrogate method in more detail. Also, Beaumont (2020) and Rao (2020) reviewed available methods to use data from a nonprobability source, the literature on combining information from probability sample and nonprobability sample, and concluded on recent approaches which are not reliable or general enough to improve population prediction and inference.

In this research, we focus on binary data which occurs in many different situations. In statistics, binary data are a type of categorical data that can only be two possible values, such as "yes" and "no", or "head" and "tail". For binary variable, which is a random variable of binary type, it follows a Bernoulli distribution. For our application on body mass index, we analyzed "obesity" and "non-obesity" categories.

The main idea of this research is to compare the performance of nine methods with different constructed survey weights, and we can use these methods for non-probability sampling after weights are estimated (e.g. quasi-randomization). In this chapter, we introduce basic terms on a sample survey, and then, discuss two methods - frequentist modeling and Bayesian modeling on survey sampling. In Chapter 2, we describe our Bayesian methodology that uses nine different types of survey weights. A simulation study is performed in Chapter 3, to make further comparisons of these nine models. In Chapter 4, we discuss an application on body mass index and compare the nine methods. Finally, Chapter 5 reviews the strengths and weaknesses of our study in more detail and provides some

future research suggestions.

1.2 Survey Sampling

As we all know, a good sample, which is a set of a population, should be represented in the sense that our interests in the population could be estimated from the sample with a known degree of accuracy. In an ideal survey, the sampled population should be the same as our target population, but in reality, the sampled population is usually much smaller than the target population. However, when some part of our target population is not in the sampled population, selection bias occurs. For example, a sample of convenience is usually biased, because the samples which are easier to select or that are more likely to respond to are often not representative of the inconvenient to select or non-responding samples.

1.2.1 Simple Probability Samples

A simple random sample is the simplest form of a probability sample. When every possible sample of the population has the same chance of being selected, this is the simple random sample.

In probability sampling, each possible sample from the population has a known probability of being chosen. We calculate $\pi_i = P(\text{Unit } i \text{ in sample})$ by adding up the probabilities of all possible samples that include unit i, and this is also called selection probability. Definitely, $\sum_{i=1}^{N} \pi_i = n$. Then, we can define the sampling weight, for any sampling methods, to be the reciprocal of the selection probability, $w_i = \frac{1}{\pi_i}$. Also, the sampling weight of i-th unit is the number of population units which can be represented by unit i.

Considered the most basic form of probability sampling, simple random sampling provides the basic theoretical concepts for the more complicated sampling. There are two ways to perform a simple random sample: with replacement or without replacement. For a simple random sample with replacement of size, n from a population of size N, each unit is randomly selected from the population with the same probability $1/N$; for a simple random

sample without replacement, the probability of selecting any sample S including n units is $1/\binom{N}{n}$.

As we mentioned before, y_i, $i = 1,\ldots,N$, is the variable of interest of the i-th unit, where $N < \infty$ is the total number of units in the population, and n is the sample size. In a simple random sampling, each unit has the same inclusion probability $\pi_i = \frac{n}{N}$; correspondingly, all sampling weights are the same with $w_i = \frac{1}{\pi_i} = \frac{N}{n}$. Note that, for a simple random sampling,

$$\sum_{i\in S} w_i = \sum_{i\in S} \frac{N}{n} = N,$$

$$\sum_{i\in S} w_i y_i = \sum_{i\in S} \frac{N}{n} y_i = \hat{t},$$

$$\frac{\sum_{i\in S} w_i y_i}{\sum_{i\in S} w_i} = \frac{\hat{t}}{N} = \hat{\bar{y}}.$$

All survey weights are the same in a simple random sampling. In other words, every unit in the sample represents the same number of units, N/n, in the population.

1.2.2 Sampling with Unequal Probabilities

Suppose a town has three supermarkets, ranging in size from 100 square feet to 1000 square feet, and we want to estimate the total amount of sales in the four stores for last week by sampling just one of the stores. Of course, one larger supermarket could have more sales than a smaller supermarket. In this situation, the probability that a store is selected should be related to its square feet, which leads to unequal survey weights. For example, let store A account for 1/15 of the total floor area of the three stores, so it is sampled with a selection probability of 1/15. For illustrative purposes, we have the following table to show the relationship between the size of the store and its selection probability π_i:

Store	Size (sqrt)	π_i	y_i (in Thousands)
A	100	$\frac{1}{15}$	12
B	400	$\frac{4}{15}$	78
C	1000	$\frac{10}{15}$	210
Total	1500	1	300

This example is the probability proportional to size (PPS) sampling, i.e. the selection probability is the proportion of the relative size of the i-th unit. In addition, if a sampling process where every unit of the population is subjected to an independent Bernoulli trial can determine whether the unit is a part of the sample or not, it calls Poisson sampling.

In a simple random sampling of size n of a population of size N, for all units i

$$w_i = \frac{N}{n}.$$

In stratified sampling, the sampling weight for item i in stratum h, in which n_h units will be sampled out of N_h total, is

$$w_{ih} = \frac{N_h}{n_h}.$$

In one stage cluster sampling with n clusters sampled out of N total, the sampling weight for cluster i is $w_i = N/n$ and for each unit j inside of it

$$w_{ij} = \frac{N}{n}.$$

In two-stage cluster sampling with n clusters sampled out of N total, the sampling weight for cluster i is $w_i = N/n$ and for each unit j inside of it, where m_i secondary sampling units are going to be sampled out of M_i total

$$w_{ij} = \frac{N}{n}\frac{M_i}{m_i}.$$

1.2.3 Non-probability Sampling Methods

Non-probability sampling is defined as a sampling mechanism in which the researchers select samples based on the subjective judgment of researchers rather than random selection. In general, it is a sampling method in which not all units of the population have an equal chance of being selected, unlike probability sampling, where every unit has the non-zero probability to be chosen. To deal with non-probability sampling, we have two main methodologies, design-based approaches and model-based approaches (Särndal et al. 1978, Gregoire 1998).

Classical survey sampling relies on design-based methods. This means that the most important source of randomness is the probability described by the sampling design to the various subsets of the finite population. In other words, the concept of sampling design plays a significant role in classical survey sampling theory, where the design could specify the randomized way in which a sample is selected from the finite population, like simple random sampling, stratified random sampling, cluster sampling, two-stage sampling, etc. The choice of design is always inspired by administrative or practical causes, but also by the desire of making full use of auxiliary information. Design-based approaches proceed by suggesting one or more suitable estimators for each given design.

However, in a world of "big data", for a large amount of available data which are easier and faster to obtain than probability samples, classical design-based approaches are hard to apply directly for making population inference, even if the data elements are selected randomly. The main cause is that the selection probabilities are missing or the selec-

tion mechanism is unknown in non-probability sampling (Chen, Li, and Wu 2020).

For inference from big data or non-probability samples, model-based methods are widely applied. Model-based approaches assume a mathematical or statistical model, which is ordinarily a linear or generalized linear model (Valliant 2000). Baker et al. (2013) summarized adjusted methods rely on models and external auxiliary information, which is provided by a task force of the American Association for Public Opinion Research (AAPOR). Buelens, Burger, and Brakel (2018) compared different model-based inference methods for non-probability samples and presented a simulation study using real-world data.

To make an inference about a finite population, Rao (2020) stated that it is possible to use a single non-probability sample only. In the presence of a relevant probability sample and a set of common auxiliary variables, we can use propensity scores to estimate survey weights and then consider the non-probability samples as regular probability samples, which is also known as quasi-randomization (Lee and Valliant 2009, Elliott and Valliant 2017). In addition, another way is fitting models on the non-probability sample and then make predictions on the response variable for units in the probability sample (Kim et al. 2018, Wang et al. 2015).

1.3 Frequentist Modeling vs. Bayesian Modeling

This section provides an account of both frequentist modeling and Bayesian modeling to inference from survey samples, focusing on descriptive parameters of a finite population.

1.3.1 Frequentist Modeling

The frequentist modeling to inference assumes that the population structure is a specified population model and that the sample holds the same model, which means there is no sample selection with respect to the assumed population model. Usually, distribution

assumptions can be avoided by focusing on variance estimation, point estimation, or associated confidence interval, as design-based inference, which causes applied models to specify the mean function and the variance function of our variable of interest (Rao et al. 2011). An important advantage of frequentist modeling is that it leads to inferences conditional on the selected sample of units.

1.3.2 Parametric Bayesian Approach

Under a specified distribution on the assumed model, it is easy to implement Bayesian inference, by given the model holds for the sample. Royall and Pfeffermann (1982) discussed Bayesian inference on the finite population with normality assumption and flat priors on the parameters of a linear regression model. To get inference on the more complicated model, we can use the Monte Carlo Markov chain (MCMC) method to simulate samples from the posterior distribution of parameters.

Musal et al. (2012) present a Bayesian framework for population utility estimation. Pfeffermann, Da Silva Moura, and Do Nascimento Silva (2006) discussed an application of the Bayesian modeling to make inferences from multilevel models under informative sampling. Nevertheless, in this situation, the multilevel sample model, which is induced by informative sampling, is more complicated and different than the corresponding population model, which means it is hard to use frequentist methods. This author shows it is efficient and convenient to use non-informative priors on the model parameters and applied MCMC methods for dealing with such complex sample models. This application presents the computational advantage provided by parametric Bayesian modeling over the corresponding frequentist modeling.

1.3.3 Non-parametric Bayesian Approach

When it comes to multipurpose surveys, the parametric Bayesian approach based on distribution assumptions is restricted because it is difficult to make valid model assump-

tions. In this situation, the non-parametric Bayesian approach is appealing to be applied.

Chipman et al. (2010) introduced a strong non-parametric Bayesian predictive tool-Bayesian Additive Regression Trees (BART)-by automatically and randomly capturing non-linear associations and high-order interactions. The idea of BART is that using the sum of trees regressions to approximate the outcome variables or our interests as an arbitrary function of predictors. Kern et al. (2016) and Wendling et al. (2018) elucidate the advantages of BART for predictive inference of population to protect against model misspecification, Rafei et al. (2021) expand the idea of double robustness such that more flexible non-parametric methods, like BART, as well as Bayesian models, could be used for prediction.

Another robust approach is taking the Dirichlet process into consideration. It has been shown that the Dirichlet process prior is useful because it makes the process more robust, and the Bayesian method helps to reduce the effect of unidentifiable prominent in non-negligible and non-response models (Nandram and Choi 2004). Furthermore, the Dirichlet process model, Dirichlet mixture model, and Dirichlet process Gaussian model can also be applied to deal with data noise (Nandram and Yin 2016a, Nandram and Yin 2016b).

Chapter 2

A Bayesian Predictive Inference Model

When a biased sample is selected from a finite population, it is hard to make inferences about the population since the nature of bias is unknown, such as when a sample is selected from a finite population with probability proportional to size (PPS).

In PPS sampling, the selection probabilities of samples are proportional to some measure of size, and this measure of size is also proportional to the characteristic of our interest. Let y_i, $i = 1, \ldots, N$, be the variable of interest of the i-th unit, where $N < \infty$ is the total number of units in the population. π_i, $i = 1, \ldots, N$, denote the selection probabilities of the entire finite population. Therefore, $\pi_i = \beta_0 + \beta_1 y_i$, $i = 1, \ldots, N$. For example, larger fish have a higher probability to be captured by net, and they do be larger in length so that when we focus on the length of fish in our pond, it is reasonable to take the selection probabilities to be linearly related to the length of fish.

Suppose $f(\underset{\sim}{y} \mid \underset{\sim}{\theta_1})$ is the probability distribution of finite population, and we draw samples with weights function $w(\underset{\sim}{y}; \underset{\sim}{\theta_1}, \underset{\sim}{\theta_2})$. Therefore, the probability distribution of samples is updated to $p(\underset{\sim}{y} \mid \underset{\sim}{\theta_1}, \underset{\sim}{\theta_2})$. In other words, the non-representative sample is observed from

$$p\left(\underset{\sim}{y} \mid \underset{\sim}{\theta_1}, \underset{\sim}{\theta_2}\right) \propto w\left(\underset{\sim}{y}; \underset{\sim}{\theta_1}, \underset{\sim}{\theta_2}\right) f\left(\underset{\sim}{y} \mid \underset{\sim}{\theta_1}\right). \tag{1}$$

In this chapter, we try to answer the question that "Can we recreate the probability sample from finite population $f(\underset{\sim}{y} \mid \underset{\sim}{\theta_1})$?". By adjusting the sample weights, we create different surrogate samples from the original finite population and compare their performance of making inferences about the original finite population via $f(\underset{\sim}{y} \mid \underset{\sim}{\theta_1})$. In the model-based analysis, PPS sampling is a special case within the Bayesian framework that

deserves special attention. To compare the different drawing methods, Poisson sampling is also considered.

2.1 Simple Random Sampling Model

Suppose a simple random sample of size n is selected from a finite population of size N, $y_s = (y_1, \ldots, y_n)'$ and $y_{ns} = (y_{n+1}, \ldots, y_N)'$ are the sampled and non-sampled units. Assume our data are binary,

$$y_1 \ldots, y_N \mid p \overset{\text{i.i.d}}{\sim} \text{Bernoulli } (p). \tag{2}$$

Under the Bayesian framework, since there is no information about p, we consider the proper but non-informative prior

$$p \sim \text{ uniform } (0,1). \tag{3}$$

According to Bayesian theorem, when we have a prior on unknown parameters $\underset{\sim}{\theta}$, and the likelihood function based on observed data $\underset{\sim}{y}$ is $f(\underset{\sim}{y}|\underset{\sim}{\theta})$, then

$$\pi(\underset{\sim}{\theta}|\underset{\sim}{y}) = \frac{f(\underset{\sim}{y}|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})}{\int f(\underset{\sim}{y}|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})d\underset{\sim}{\theta}},$$

where $\pi(\underset{\sim}{\theta}|\underset{\sim}{y})$ is posterior distribution on unknown parameters $\underset{\sim}{\theta}$, and can be written as well as

$$\pi(\underset{\sim}{\theta}|\underset{\sim}{y}) \propto f(\underset{\sim}{y}|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta}). \tag{4}$$

Therefore, the posterior distribution is

$$p \mid \underset{\sim}{y_s} \sim \text{ Beta } (t_s + 1, n - t_s + 1), \tag{5}$$

where $t_s = \sum_{i=1}^{n} y_i$, then it is reasonable to make inference about p via the posterior distri-

bution.

Let $T = \sum_{i=1}^{N} y_i$, which is the finite population total, and $T_{ns} = \sum_{i=n+1}^{N} y_i$ is for non-sampled units. When it comes to finite population proportion, it can be calculated as $P = T/N = f\bar{y}_s + (1-f)\bar{y}_{ns}$, where $f = \frac{n}{N}$ is the weight, and $\bar{y}_s = \sum_{i=1}^{n} y_i/n$, $\bar{y}_{ns} = \sum_{i=n+1}^{N} y_i/(N-n)$ are two parts with respect to sampled units and non-sampled units. In this way, to make prediction of non-sampled units, Bayesian predictive inference formula is useful,

$$f\left(T_{ns} \mid \underset{\sim}{y_s}\right) = \int_0^1 f\left(T_{ns} \mid p\right) \pi\left(p \mid \underset{\sim}{y_s}\right) dp, \tag{6}$$

where $\pi(p \mid \underset{\sim}{y_s})$ is the posterior distribution of p, given by Equation 5, and it is easy to show that $T_{ns} \mid p \sim$ Binomial $(N-n, p)$. Then, through Equation 6 we can combine observed sampling information and non-sampling information to make inference about our interests of population.

2.2 Sampling Models

In a biased sample, Equation 5 and Equation 6 are no longer satisfied. In this situation, the surrogate sampling method allows us to obtain a surrogate random sample from a population, and then make an inference of our interests of the population. In most situations, weights as well as covariates are known, but not the response y, in the probability samples. To handle this problem of the biased estimate, we discuss nine survey weights models.

In general, in this section, we discuss how to obtain $f(\underset{\sim}{y} \mid \underset{\sim}{\theta_1})$ in Equation 1 by adjusting the sample weights, then generate different surrogate samples from original finite population to make inference of our interests of the population.

2.2.1 Survey Weights

Let y_i be a vector of response for a unit i; x_i be a vector of observed covariates for a unit i; u_i be a vector of unobserved covariates for a unit i; z_i be an indicator variable for unit i being selected ($z_i = 1$ if the unit i was in the probability sample and $z_i = 0$ if the unit i was in the non-probability sample; W_i be the survey weight for unit i; π_i is the selection probability that the unit i was selected given all its features, under the assumption that the selection probability only depends on observed covariates,

$$\pi_i = P(z_i = 1 \mid x_i, u_i, y_i) = P(z_i = 1 \mid x_i), \quad i = 1, \ldots, N.$$

When a sampling plan is implemented in a finite population of size N to draw a size n sample, with given selection probabilities $\pi_1, \ldots, \pi_N$ to each selected unit, since $W_i = 1/\pi_i$, $i = 1, \ldots, N$, the unbiased estimator of population total and population size are

$$\begin{aligned} \hat{T} &= \textstyle\sum_{i=1}^{n} y_i W_i, \\ \hat{N} &= \textstyle\sum_{i=1}^{n} W_i. \end{aligned} \tag{7}$$

Potthoff, Woodbury, and Manton (1992) mentioned the effective sample size,

$$n_e = \frac{(\sum_{i=1}^{n} W_i)^2}{\sum_{j=1}^{n} W_j^2}.$$

The effective sample size indicates the degree to which the variance increases due to the unequal weight. Then, the adjustment weight required to eliminate the bias introduced by the weight is

$$w_i = \frac{n_e W_i}{\sum_{j=1}^{n} W_j}, \quad i = 1, \ldots, n. \tag{8}$$

Here, we use capital W for original survey weights, and small w for adjusted survey weights. The effective sample size n_e has some interesting properties. For example, by calculation, we get $n_e = \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i^2$; When W_i are almost equal, $n_e \approx n$.

If replace n_e to n in Equation 8, we obtain a new adjusted weight, adjusted standardized weight w^*,

$$w_i^* = \frac{nW_i}{\sum_{i=1}^{n} W_i}, \quad i = 1, \ldots, n, \tag{9}$$

and this adjusted standardized weight w^* satisfied $n = \sum_{i=1}^{n} w_i^* = \sum_{i=1}^{n} w_i^{*2}$.

To improve statistical efficiency and increase the robustness of statistical inferences, Winsorization is an effective way to deal with outliers (Rao 1966, Basu 1971, Haziza, Beaumont, et al. 2017). Outliers here are defined as observations fall above $Q_3 + 1.5(Q_3 - Q_1)$, where Q_1=1st quartile, Q_3=3rd quartile. Let W^* be weights after trimming and W_0 be the threshold, which value is $Q_3 + 1.5(Q_3 - Q_1)$,

$$W_i^* = \begin{cases} W_0, & W_i \geq W_0 \\ \\ aW_i, & W_i < W_0 \end{cases}, \tag{10}$$

where a is a rescaling parameter such that $\sum_{i=1}^{n} W_i^* = \sum_{i=1}^{n} W_i$.

To sum up, we introduced original survey weights W, adjusted survey weights w, adjusted standardized survey weights w^*, and trimmed survey weights W^*.

2.2.2 Posterior Distributions

In this part, different survey samples are incorporated with Equation 5 based on Equation 7 to develop Bayesian posterior distributions, and we denote them using nine cases.

- Case A: The simplest model if we ignore the survey weights and use responses of non-probability sampling to make inference of interests of population. Since $y_1 \ldots, y_N \mid p \overset{\text{i.i.d}}{\sim}$ Bernoulli (p) and $p \sim$ uniform $(0,1)$, the posterior distribution is,

$$\theta \mid \underset{\sim}{y} \sim \text{Beta}\left(\sum_{i=1}^{n} y_i + 1, n - \sum_{i=1}^{n} y_i + 1\right). \tag{11}$$

  As we discussed in the previous section, the model ignoring survey weights would be biased, and by comparing other cases with this one, it is supposed to show how biased samples influence our inference of population. In other words, inference of population is made according to $p(\underset{\sim}{y} \mid \underset{\sim}{\theta_1}, \underset{\sim}{\theta_2})$ in Equation 1, which is obviously biased.

In following cases, our goal is to obtain $f(\underset{\sim}{y} \mid \underset{\sim}{\theta_1})$ by incorporating different survey weights with Equation 1. Case B and case C are related to original survey weights W.

- Case B: Replace sample total with estimator of population total in Equation 5 by using original survey weights,

$$\theta \mid \underset{\sim}{y} \sim \text{Beta}\left(\sum_{i=1}^{n} y_i W_i + 1, \sum_{i=1}^{n} (1 - y_i) W_i + 1\right). \tag{12}$$

- Case C: Consider the normalized likelihood with original survey weights,

$$f_C(\theta \mid \underset{\sim}{y}) \propto \frac{\theta^{\Sigma_{i=1}^{n} y_i W_i} (1-\theta)^{\Sigma_{i=1}^{n} (1-y_i) W_i}}{\prod_{i=1}^{n} [\theta^{W_i} + (1-\theta)^{W_i}]}. \tag{13}$$

  Here, we used the normalized likelihood to update the posterior distribution according to Equation 4.

The following models are generated by using adjusted survey weights w (case D and case E), adjusted standardized survey weights w^* (case F and case G), and trimmed survey weights W^* (case H and case I).

- Case D: Beta distribution with adjusted survey weights w,

$$\theta \mid \underset{\sim}{y} \sim \text{Beta}\left(\sum_{i=1}^{n} y_i w_i + 1, \sum_{i=1}^{n} (1 - y_i) w_i + 1\right). \tag{14}$$

- Case E: The normalized likelihood with adjusted survey weights w,

$$f_E(\theta \mid \underset{\sim}{y}) \propto \frac{\theta^{\sum_{i=1}^{n} y_i w_i} (1-\theta)^{\sum_{i=1}^{n} (1-y_i) w_i}}{\prod_{i=1}^{n} [\theta^{w_i} + (1-\theta)^{w_i}]}. \tag{15}$$

- Case F: Beta distribution with adjusted standardized survey weights w^*,

$$\theta \mid \underset{\sim}{y} \sim \text{Beta}\left(\sum_{i=1}^{n} y_i w_i^* + 1, \sum_{i=1}^{n} (1 - y_i) w_i^* + 1\right). \tag{16}$$

- Case G: The normalized likelihood with adjusted standardized survey weights w^*,

$$f_G(\theta \mid \underset{\sim}{y}) \propto \frac{\theta^{\sum_{i=1}^{n} y_i w_i^*} (1-\theta)^{\sum_{i=1}^{n} (1-y_i) w_i^*}}{\prod_{i=1}^{n} \left[\theta^{w_i^*} + (1-\theta)^{w_i^*}\right]}. \tag{17}$$

- Case H: Beta distribution with trimmed survey weights W^*,

$$\theta \mid \underset{\sim}{y} \sim \text{Beta}\left(\sum_{i=1}^{n} y_i W_i^* + 1, \sum_{i=1}^{n} (1 - y_i) W_i^* + 1\right). \tag{18}$$

- Case I: The normalized likelihood with trimmed survey weights W^*,

$$f_I(\theta \mid \underset{\sim}{y}) \propto \frac{\theta^{\sum_{i=1}^{n} y_i W_i^*} (1-\theta)^{\sum_{i=1}^{n} (1-y_i) W_i^*}}{\prod_{i=1}^{n} \left[\theta^{W_i^*} + (1-\theta)^{W_i^*}\right]}. \tag{19}$$

Chapter 3

Simulation Study

In this chapter, one simulation is studied to assess the performance of our nine different models.

The design of our simulation is inspired by the one implemented in Nandram (2007). The samples are given a random selection mechanism with unequal probabilities. But in the simulation chapter, to generate probability samples, it is assumed that at the stage of analysis, the selection probabilities are known, and our goal is to adjust for the selection bias by using a probability sample whose weights are known. We conduct the simulation under nine Bayesian models.

Consider a finite population of size N, and the sample units $y_1,\ldots,y_N$ are drawn with probability proportional to measures of size $x_1,\ldots,x_N$, which should be non-negative. Here, x is auxiliary variable and z is a latent variable, generated as follows:

$$x_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha,\beta), \tag{20}$$

$$z_i \mid x_i \overset{\text{iid}}{\sim} N\left(\frac{\rho}{\sigma_x}(x_i-\mu),1-\rho^2\right), \quad i=1\ldots N, \tag{21}$$

where ρ is the correlation coefficient, $\mu=\frac{\alpha}{\beta}$ and $\sigma_x^2=\frac{\alpha}{\beta^2}$.

Then, we use the latent variable to generate binary responses,

$$y_i=\begin{cases} 1, & z_i\geq 0 \\ \\ 0, & z_i<0 \end{cases}, \tag{22}$$

where $i = 1 \ldots N$.

In this simulation, assume that our interest is the population proportion, $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$. In a biased sample, the sampled values are taken with unequal selection probabilities which depend on the characteristic y. For the probability proportional to size sampling, construct selection probabilities,

$$\pi_i = \frac{nx_i}{\sum_{i=1}^{N} x_i}, \quad i - 1, \ldots, N. \tag{23}$$

For the Poisson sampling, samples are selected as follows,

$$I_i \sim Bernoulli(\pi_i), \quad i = 1, \ldots, N,$$

when $I = 1$, individual y is selected and when $I = 0$, individual y is not selected from population. Letting $n_0 = \sum_{i=1}^{N} I_i$ denote the size of sample in Poisson sampling, we have $E(n_0) = \sum_{i=1}^{N} E(I_i) = n$, and $Var(n_0) = \sum_{i=1}^{N} \pi_i(1 - \pi_i)$, so $n_0 \approx n$.

Now, we can perform the simulation study to access the estimators of the finite population under probability proportional to size (PPS) and Poisson sampling with respect to measure of size $x_i, i - 1, \ldots, N$, and the effective size of samples should be equal to or around $n = 100$ according to different sampling method. Keeping the population size fixed at $N = 1000$ and $\beta = 1$, we can generate $K = 10000$ datasets at $\rho = \{0.2, 0.5, 0.8\}$ and $\alpha = \{2, 5, 15\}$, which means there are nine design points for each posterior distribution.

To evaluate the repeated surrogate sampling properties of our nine models, posterior mean (PM), posterior standard deviation (PSD), relative bias (RB), posterior root mean squared error (PRMSE) and the proportion of the 100 95% credible intervals containing θ (PCI) are calculated as below:

$$PM(\widehat{\bar{y}}) = \frac{1}{K}\sum_{k=1}^{K}\widehat{\bar{y}}^{(k)}, \tag{24}$$

$$PSD(\widehat{\bar{y}}) = \frac{1}{K}\sum_{k=1}^{K}\text{var}\left(\widehat{\bar{y}}^{(k)}\right), \tag{25}$$

$$RB(\widehat{\bar{y}}) = \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\bar{y}}^{(k)} - \bar{y}\right)/\bar{y}, \tag{26}$$

$$PRMSE(\widehat{\bar{y}}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\bar{y}}^{(k)} - \bar{y}\right)^2}/\bar{y}, \tag{27}$$

$$PCI(\widehat{\bar{y}}) = \frac{1}{K}\sum_{k=1}^{K} I\left(\left|\widehat{\bar{y}}^{(k)} - \bar{y}\right| < z_{0.975}\sqrt{\text{var}\left(\widehat{\bar{y}}^{(k)}\right)}\right), \tag{28}$$

where $\widehat{\bar{y}}$ denotes the posterior surrogate mean from iteration k, $\bar{y}$ is the finite population true mean, and var$(.)$ represents the variance estimate of the posterior mean based on the surrogate sample.

For the two sampling processes in Table 1 and Table 2, we can see there is no significant difference in posterior mean for each posterior distributions and $\hat{n}$. Also, the effective sample size n_e of Poisson sampling are almost 100, satisfied that $E(n_e) = n$, and $Var(n_e)$ is small.

Keeping ρ fixed and increasing α, we can find the eight posterior distributions using survey weights performed better than A (posterior distribution without weights). The PMs of model A, C, I are closer to the true mean with increased α, but PMs of distribution B, D are further. On the other side, the distances between PM and true mean of nine posterior distributions are greater when ρ is increasing and α is fixed.

As for comparisons of the PM, there is no significant difference between PPS and Poisson sampling, and the effective size of Poisson sampling is around n in Table 3 and Table 4. With ρ fixed and α increasing, PSDs of model D, E (posterior distributions with standardized weights) become smaller but others get greater. Also, if α is fixed and ρ is

Table 1

PPS: Comparisons of the posterior mean (PM) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	θ	$\hat{n}$
0.2	2	0.5472	0.4996	0.5001	0.4996	0.4993	0.4998	0.5005	0.5095	0.5010	0.4961	58.97
	5	0.5371	0.5009	0.5001	0.5007	0.5006	0.5009	0.5011	0.5051	0.5004	0.5033	80.62
	15	0.5109	0.4910	0.4992	0.4913	0.4910	0.4911	0.4917	0.4920	0.4994	0.4992	93.39
0.5	2	0.6063	0.4769	0.4985	0.4776	0.4763	0.4774	0.4859	0.5039	0.5004	0.4900	57.91
	5	0.5705	0.4828	0.4987	0.4830	0.4826	0.4831	0.4860	0.4927	0.4994	0.4932	79.66
	15	0.5490	0.4994	0.5000	0.4990	0.4995	0.4993	0.4992	0.5015	0.5002	0.4973	93.49
0.8	2	0.6416	0.4351	0.4963	0.4377	0.4285	0.4361	0.4610	0.4833	0.4990	0.4562	56.90
	5	0.5921	0.4551	0.4963	0.4561	0.4551	0.4559	0.4637	0.4687	0.4973	0.4711	80.77
	15	0.5676	0.4877	0.4989	0.4877	0.4879	0.4880	0.4890	0.4905	0.4992	0.4833	93.43

increasing, PSDs of model D and E are greater but others are smaller. Further, based on the type of survey weights, we compare models in pairs and we can conclude that PSDs of models ignoring denominator are larger than PSDs of models considering denominator, except for model D and E, where D (without denominator) is smaller than E (with denominator). D and E have larger PSDs than others and model B and C (with raw weights), H and I (with trimmed weights) have similar and smaller PSDs.

When ρ is fixed, RBs of almost all models are smaller with increasing α but RBs of model G seem to have an increasing trend in Table 5 and Table 6. When α is fixed and ρ is increasing, RBs of model A and I are increasing, too. But RBs of the others are smaller than greater. Also, the tables show that model C and I have smaller RBs.

According to Table 7 and Table 8, these two tables indicate that model C and model I have smaller PRMSEs than the other models. Keeping ρ constant and changing α from 2 to 15, we can figure out that PRMSEs of almost all models are negative related to α, except

Table 2

Poisson sampling: Comparisons of the posterior mean (PM) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	θ	$\hat{n}$	n_e
0.2	2	0.5537	0.4961	0.4999	0.4962	0.4941	0.4961	0.4989	0.5096	0.5009	0.4975	59.87	100.32
	5	0.5315	0.4980	0.4998	0.4981	0.4983	0.4981	0.4980	0.4998	0.5000	0.5019	81.00	100.35
	15	0.5256	0.5059	0.5007	0.5056	0.5055	0.5060	0.5056	0.5068	0.5006	0.5012	92.74	99.12
0.5	2	0.6088	0.4862	0.4991	0.4868	0.4851	0.4865	0.4919	0.5139	0.5012	0.4888	58.83	99.84
	5	0.5768	0.4931	0.4994	0.4935	0.4930	0.4932	0.4949	0.5010	0.5002	0.4918	80.96	99.55
	15	0.5558	0.5068	0.5007	0.5066	0.5065	0.5068	0.5063	0.5090	0.5010	0.4972	93.80	100.36
0.8	2	0.6599	0.4510	0.4969	0.4529	0.4463	0.4519	0.4713	0.5017	0.5003	0.4549	57.55	100.26
	5	0.6149	0.4784	0.4983	0.4788	0.4782	0.4787	0.4826	0.4942	0.4997	0.4697	81.57	101.45
	15	0.5547	0.4743	0.4976	0.4747	0.4748	0.4749	0.4760	0.4777	0.4979	0.4840	95.17	102.01

model G, PRMSEs of which get larger when α increases. If the α is constant, PRMSEs of A tend to be greater with increasing ρ, but there is no significant pattern for other models.

The Table 9 and Table 10 indicate that model D, E and G have the highest proportion that true mean θ is included in the 95% credible interval. However, it is surprising that when $\rho = 0.8$ and $\alpha = 2$, the PCIs of some models are too small. For instance, only one 95% credible interval of model A in Poisson sampling contains the true mean from the 100 datasets.

To see more details, we draw the histogram of surrogate posterior means, true mean, and 95% credible intervals under nine models for one dataset. Here we can see from Figure 1, the true mean is 0.454 and $\rho = 0.8$, $\alpha = 2$. For model A, if we ignore the survey weights and only consider the posterior distribution of our interest, it is possible to get left-side skewed distribution which is not good for inference. Model B, model C, model H, and model I used the raw weights and trimmed weights, but didn't be standardized, so the distributions are sharper than others, which used the standardized weights. For model D,

Table 3

PPS: Comparisons of the posterior standard deviation (PSD) using nine posterior distributions of the finite population by ρ *and* α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$
0.2	2	0.0511	0.0221	0.0166	0.0658	0.0691	0.0511	0.0434	0.0222	0.0167	58.97
	5	0.0515	0.0221	0.0167	0.0566	0.0571	0.0513	0.0475	0.0221	0.0167	80.62
	15	0.0514	0.0223	0.0168	0.0530	0.0530	0.0516	0.0501	0.0223	0.0167	93.39
0.5	2	0.0503	0.0220	0.0166	0.0663	0.0691	0.0511	0.0429	0.0220	0.0166	57.91
	5	0.0510	0.0222	0.0167	0.0569	0.0577	0.0513	0.0474	0.0221	0.0167	79.66
	15	0.0513	0.0222	0.0168	0.0529	0.0532	0.0514	0.0501	0.0222	0.0168	93.49
0.8	2	0.0492	0.0218	0.0166	0.0674	0.0716	0.0510	0.0435	0.0219	0.0167	56.90
	5	0.0507	0.0221	0.0167	0.0564	0.0571	0.0514	0.0478	0.0222	0.0167	80.77
	15	0.0510	0.0222	0.0168	0.0528	0.0531	0.0513	0.0502	0.0222	0.0169	93.43

model E, model F, and model G, after survey weights standardized, posterior distributions are flattened and it causes the higher probability to cover the real population mean. This can also be speculated from their small proportion values (Table 9 and Table 10).

Table 4

Poisson sampling: Comparisons of the posterior standard deviation (PSD) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$
0.2	2	0.0511	0.0221	0.0166	0.0658	0.0691	0.0511	0.0434	0.0222	0.0167	58.97
	5	0.0515	0.0221	0.0167	0.0566	0.0571	0.0513	0.0475	0.0221	0.0167	80.62
	15	0.0514	0.0223	0.0168	0.0530	0.0530	0.0516	0.0501	0.0223	0.0167	93.39
0.5	2	0.0503	0.0220	0.0166	0.0663	0.0691	0.0511	0.0429	0.0220	0.0166	57.91
	5	0.0510	0.0222	0.0167	0.0569	0.0577	0.0513	0.0474	0.0221	0.0167	79.66
	15	0.0513	0.0222	0.0168	0.0529	0.0532	0.0514	0.0501	0.0222	0.0168	93.49
0.8	2	0.0492	0.0218	0.0166	0.0674	0.0716	0.0510	0.0435	0.0219	0.0167	56.90
	5	0.0507	0.0221	0.0167	0.0564	0.0571	0.0514	0.0478	0.0222	0.0167	80.77
	15	0.0510	0.0222	0.0168	0.0528	0.0531	0.0513	0.0502	0.0222	0.0169	93.43

Table 5

PPS: Comparisons of the relative bias (RB) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$
0.2	2	0.1213	0.1047	0.0283	0.1005	0.1114	0.1019	0.0659	0.0980	0.0288	58.97
	5	0.0920	0.0898	0.0240	0.0884	0.0900	0.0888	0.0732	0.0875	0.0241	80.62
	15	0.0775	0.0790	0.0209	0.0775	0.0787	0.0778	0.0734	0.0786	0.0210	93.39
0.5	2	0.2376	0.1028	0.0281	0.0989	0.1088	0.1012	0.0620	0.0852	0.0300	57.91
	5	0.1590	0.0812	0.0243	0.0802	0.0805	0.0795	0.0657	0.0787	0.0248	79.66
	15	0.1204	0.0795	0.0266	0.0778	0.0780	0.0777	0.0730	0.0798	0.0266	93.49
0.8	2	0.4073	0.1266	0.0898	0.1208	0.1412	0.1238	0.0760	0.1286	0.0954	56.90
	5	0.2570	0.0759	0.0557	0.0737	0.0747	0.0739	0.0577	0.0725	0.0577	80.77
	15	0.1762	0.0841	0.0394	0.0820	0.0832	0.0829	0.0785	0.0865	0.0397	93.43

Table 6

Poisson sampling: Comparisons of the relative bias (RB) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$	n_e
0.2	2	0.1230	0.1129	0.0244	0.1085	0.1202	0.1112	0.0711	0.0952	0.0244	59.87	100.32
	5	0.0970	0.0978	0.0274	0.0955	0.0982	0.0967	0.0815	0.0934	0.0272	81.00	100.35
	15	0.0952	0.0880	0.0242	0.0857	0.0863	0.0865	0.0820	0.0888	0.0243	92.74	99.12
0.5	2	0.2458	0.0914	0.0307	0.0878	0.0954	0.0892	0.0581	0.0897	0.0327	58.83	99.84
	5	0.1752	0.0817	0.0268	0.0796	0.0814	0.0796	0.0667	0.0804	0.0273	80.96	99.55
	15	0.1240	0.0813	0.0266	0.0799	0.0798	0.0800	0.0756	0.0820	0.0268	93.80	100.36
0.8	2	0.4505	0.1053	0.0935	0.1011	0.1151	0.1034	0.0663	0.1234	0.1007	57.55	100.26
	5	0.3092	0.0859	0.0632	0.0846	0.0853	0.0839	0.0727	0.0976	0.0658	81.57	101.45
	15	0.1518	0.0747	0.0345	0.0734	0.0737	0.0735	0.0687	0.0750	0.0349	95.17	102.01

Table 7

PPS: Comparisons of the posterior root mean square error (PRMSE) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$
0.2	2	0.0833	0.0592	0.0226	0.0874	0.0938	0.0768	0.0573	0.0558	0.0230	58.97
	5	0.0740	0.0527	0.0216	0.0759	0.0767	0.0718	0.0633	0.0515	0.0217	80.62
	15	0.0674	0.0477	0.0210	0.0690	0.0695	0.0681	0.0653	0.0475	0.0209	93.39
0.5	2	0.1281	0.0575	0.0227	0.0867	0.0922	0.0759	0.0554	0.0500	0.0234	57.91
	5	0.0968	0.0482	0.0216	0.0731	0.0738	0.0683	0.0607	0.0475	0.0218	79.66
	15	0.0824	0.0477	0.0224	0.0690	0.0694	0.0679	0.0652	0.0479	0.0224	93.49
0.8	2	0.1925	0.0638	0.0442	0.0910	0.1009	0.0801	0.0581	0.0648	0.0466	56.90
	5	0.1322	0.0448	0.0317	0.0700	0.0710	0.0660	0.0580	0.0432	0.0324	80.77
	15	0.1025	0.0483	0.0268	0.0690	0.0697	0.0681	0.0657	0.0491	0.0269	93.43

Table 8

Poisson sampling: Comparisons of the posterior root mean square error (PRMSE) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$	n_e
0.2	2	0.0846	0.0630	0.0218	0.0901	0.0969	0.0805	0.0595	0.0553	0.0218	59.87	100.32
	5	0.0753	0.0564	0.0230	0.0786	0.0801	0.0751	0.0667	0.0544	0.0229	81.00	100.35
	15	0.0750	0.0519	0.0221	0.0724	0.0728	0.0714	0.0689	0.0522	0.0221	92.74	99.12
0.5	2	0.1315	0.0524	0.0236	0.0827	0.0872	0.0716	0.0542	0.0514	0.0243	58.83	99.84
	5	0.1029	0.0487	0.0225	0.0726	0.0739	0.0688	0.0612	0.0480	0.0226	80.96	99.55
	15	0.0845	0.0483	0.0225	0.0694	0.0698	0.0685	0.0658	0.0486	0.0226	93.80	100.36
0.8	2	0.2111	0.0550	0.0457	0.0846	0.0917	0.0731	0.0561	0.0635	0.0486	57.55	100.26
	5	0.1546	0.0488	0.0345	0.0730	0.0738	0.0686	0.0616	0.0535	0.0355	81.57	101.45
	15	0.0929	0.0449	0.0247	0.0671	0.0673	0.0656	0.0631	0.0450	0.0249	95.17	102.01

Table 9

PPS: Comparisons of the proportion of the 100 95% credible intervals containing θ (PCI) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$
0.2	2	0.85	0.46	0.94	0.97	0.96	0.87	0.96	0.51	0.92	58.97
	5	0.9	0.51	0.98	0.96	0.96	0.96	0.96	0.58	0.98	80.62
	15	0.98	0.58	0.98	0.97	0.99	0.97	0.98	0.61	0.97	93.39
0.5	2	0.32	0.49	0.94	0.98	0.97	0.91	0.96	0.63	0.93	57.91
	5	0.7	0.62	0.97	0.97	0.98	0.96	0.98	0.58	0.97	79.66
	15	0.9	0.62	0.98	0.97	0.97	0.96	0.97	0.6	0.98	93.49
0.8	2	0.04	0.48	0.32	0.99	0.96	0.83	0.97	0.4	0.26	56.90
	5	0.27	0.68	0.67	0.99	0.98	0.96	0.99	0.7	0.65	80.77
	15	0.58	0.61	0.84	0.97	0.96	0.96	0.97	0.59	0.85	93.43

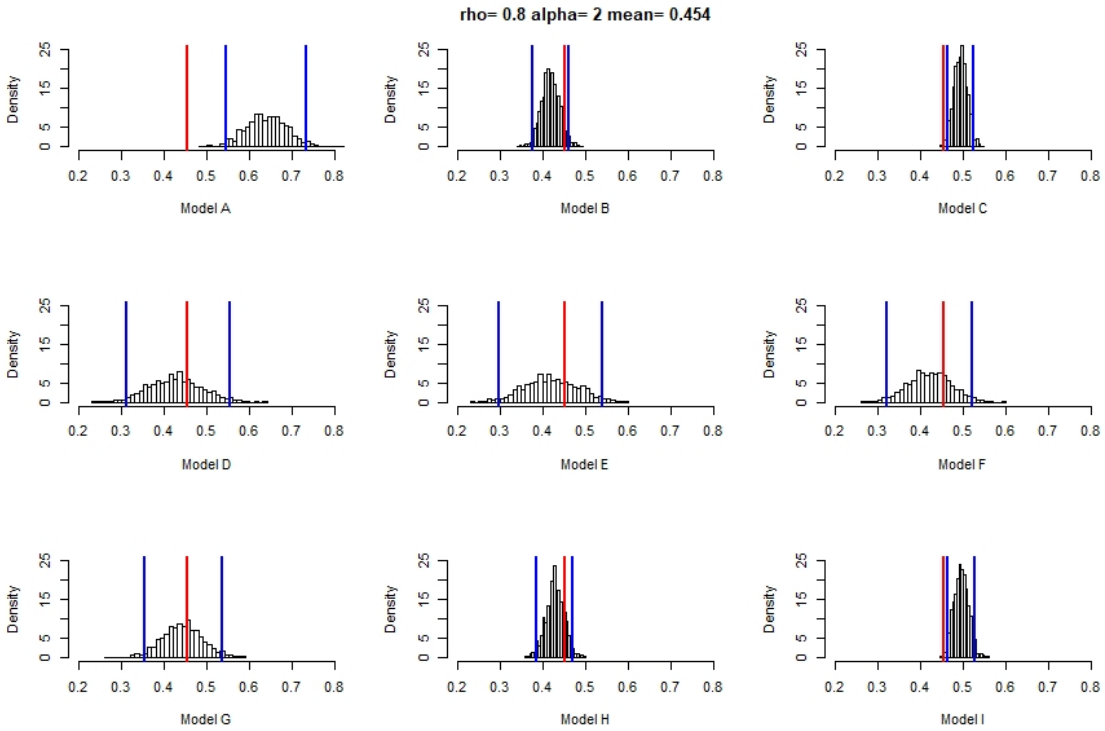
Table 10

Poisson sampling: Comparisons of the proportion of the 100 95% credible intervals containing θ (PCI) using nine posterior distributions of the finite population by ρ and α

ρ	α	A	B	C	D	E	F	G	H	I	$\hat{n}$	n_e
0.2	2	0.81	0.47	0.97	0.96	0.96	0.86	0.96	0.53	0.99	59.87	100.32
	5	0.91	0.52	0.95	0.92	0.91	0.9	0.92	0.53	0.97	81.00	100.35
	15	0.93	0.56	0.97	0.95	0.94	0.94	0.94	0.53	0.95	92.74	99.12
0.5	2	0.3	0.58	0.92	0.99	0.98	0.9	0.99	0.59	0.91	58.83	99.84
	5	0.66	0.59	0.96	0.98	0.98	0.96	0.98	0.62	0.95	80.96	99.55
	15	0.79	0.61	0.94	0.99	0.99	0.97	0.99	0.61	0.95	93.80	100.36
0.8	2	0.01	0.51	0.3	0.97	0.97	0.91	0.94	0.49	0.18	57.55	100.26
	5	0.18	0.63	0.59	0.98	0.98	0.98	0.99	0.55	0.58	81.57	101.45
	15	0.73	0.66	0.91	0.97	0.98	0.98	0.98	0.67	0.91	95.17	102.01

Figure 1

One simulated data: the histogram of surrogate posterior means, true mean (red line), and 95% credible intervals (blue lines) under the nine models.



Chapter 4

Application on Body Mass Index

In this chapter, we apply our models to the Body Mass Index (BMI) from NHANES III (Nandram and Choi 2005, Nandram and Choi 2010). In these datasets, raw sample weights for each county are given. Therefore, we can apply our nine models to these six counties as well as the whole state.

We use six counties, California, from NHANES III. The datasets contain *age*, *race* and *sex* as observed covariates, where *age* is collected as integers from 20 to 90; *race* uses $\{0, 1\}$ to denote Hispanic and non-Hispanic; *sex* is represented by 0 for male and 1 for female. Body mass index (BMI) is a simple index of weight-for-height that is commonly used to classify overweight and obesity in adults. It is defined as a person's weight in kilograms divided by the square of his height in meters (kg/m2). World Health Organization defined that obesity as a BMI greater than or equal to 30. In this dataset, we focus on obesity, which means $y_i = I(\text{BMI}_i \geq 30)$, $i = 1, \ldots, n$. Here, our sample size is $n = 1867$, including six counties.

From Table 11, model H and I (using the trimmed weights) have stable PMs under different counties. To improve the efficiency of point estimators, the excess weights are redistributed below the threshold, which means the restricted weights have a great effect on the PMs, and also improve the robustness of estimators. In addition, from Table 12, model H and I also have the smaller PSDs as we expected. By trimming the weights, we are supposed to get a smaller mean square error than that of other used estimators. Also, the posterior means via models with survey weights are more likely to be smaller than the unweighted model. As we see in the simulation study, the posterior distribution of model B, model C, model H, and model I are sharper than others and have smaller posterior standard deviations. In addition, if we consider the normalized likelihood, the PMs would be smaller

Table 11

Posterior mean (PM) of nine models for BMI data by different areas

	A	B	C	D	E	F	G	H	I	$\hat{n}$	n
All	0.2277	0.1917	0.1913	0.1931	0.0911	0.1925	0.1998	0.2197	0.2205	498.43	1867
1	0.2537	0.2198	0.2188	0.2255	0.1876	0.2233	0.2140	0.2300	0.2279	76.31	164
19	0.2474	0.1683	0.1661	0.1802	0.0965	0.1697	0.1726	0.2567	0.2555	49.88	176
37	0.2279	0.2353	0.2337	0.2382	0.1853	0.2356	0.2131	0.2345	0.2332	180.26	795
71	0.2367	0.2082	0.2069	0.2235	0.1550	0.2108	0.2048	0.2336	0.2313	33.38	125
73	0.2506	0.1585	0.1570	0.1711	0.1281	0.1628	0.1791	0.2333	0.2328	40.56	141
85	0.2149	0.1504	0.1490	0.1629	0.0649	0.1557	0.1637	0.1713	0.1701	55.82	128

than just using the Beta posterior distribution.

Table 12

Posterior standard deviation (PSD) of nine models for BMI data by different areas

	A	B	C	D	E	F	G	H	I	$\hat{n}$	n
All	0.0162	0.0125	0.0122	0.0219	0.0178	0.0151	0.0190	0.0129	0.0134	498.43	1867
1	0.0364	0.0134	0.0130	0.0470	0.0503	0.0341	0.0391	0.0135	0.0131	76.31	164
19	0.0346	0.0115	0.0123	0.0565	0.0529	0.0302	0.0490	0.0142	0.0142	49.88	176
37	0.0191	0.0137	0.0134	0.0336	0.0389	0.0201	0.0207	0.0139	0.0134	180.26	795
71	0.0393	0.0133	0.0127	0.0709	0.0746	0.0384	0.0443	0.0133	0.0136	33.38	125
73	0.0386	0.0114	0.0118	0.0561	0.0631	0.0328	0.0462	0.0142	0.0136	40.56	141
85	0.0390	0.0114	0.0112	0.0472	0.0355	0.0336	0.0449	0.0128	0.0115	55.82	128

Chapter 5

Concluding Remarks

The discussion is motivated by the desire to make predictions and inferences about our finite population from biased samples by generating surrogate samples.

According to our simulation study, the performance of these nine models is different and depends on the correlation between covariates and responses. In the BMI dataset analysis, these nine models are able to deal with the dataset that contain extreme survey weights and make proper inferences of the population, which shows the advantage of the Bayesian framework.

Also, we can consider standardizing the trimmed weights to make the posterior distribution more flatten. Table 13 compares model H, model I with updated models, respectively. After we standardized the trimmed survey weights, the posterior mean is slightly larger, and the posterior standard deviation increased about twice as much because of the flattened distribution.

In future work, it is reasonable to consider how to incorporate probability samples with non-probability samples to make full use of known information. Once the survey weights are obtained, they can be incorporated into a Bayesian model .

For the normality assumption on the dataset, the non-parametric methodology is helpful to relax model assumptions and can be more flexible on prediction and inference, and we can extend our models with some machine learning approaches. In particular, it is reasonable to combine our working, bias adjustment by using different survey weights on non-probability sampling, with the Dirichlet process mixture model.

Table 13

Posterior mean (PM) and posterior standard deviation (PSD) of H, I models and updated models for BMI data by different areas

	PM_H	PM_I	PSD_H	PSD_I	PM_{H^*}	PM_{I^*}	PSD_{H^*}	PSD_{I^*}	n
All	0.2197	0.2205	0.0129	0.0134	0.2197	0.2760	0.0162	0.0182	1867
1	0.2300	0.2279	0.0135	0.0131	0.2329	0.3297	0.0339	0.0365	164
19	0.2567	0.2555	0.0142	0.0142	0.2606	0.2759	0.0361	0.0362	176
37	0.2345	0.2332	0.0139	0.0134	0.2346	0.2780	0.0206	0.0209	795
71	0.2336	0.2313	0.0133	0.0136	0.2371	0.2781	0.0397	0.0416	125
73	0.2333	0.2328	0.0142	0.0136	0.2387	0.2551	0.0379	0.0412	141
85	0.1713	0.1701	0.0128	0.0115	0.1775	0.2327	0.0350	0.0419	128

References

Baker, Reg, J Michael Brick, Nancy A Bates, Mike Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau. 2013. "Summary report of the AAPOR task force on non-probability sampling." *Journal of survey statistics and methodology* 1 (2): 90–143.

Basu, D. 1971. *An essay on the logical foundations of survey sampling, Part I. Foundations of Statistical Inferences, VP Godambe and DA Sprott.*

Beaumont, Jean-Francois. 2020. "Are probability surveys bound to disappear for the production of official statistics?" *Survey Methodology* 46 (1): 1–28.

Buelens, Bart, Joep Burger, and Jan A van den Brakel. 2018. "Comparing inference methods for non-probability samples." *International Statistical Review* 86 (2): 322–343.

Chen, Yilin, Pengfei Li, and Changbao Wu. 2020. "Doubly robust inference with nonprobability survey samples." *Journal of the American Statistical Association* 115 (532): 2011–2021.

Chipman, Hugh A, et al. 2010. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics* 4 (1): 266–298.

Elliott, Michael R, and Richard Valliant. 2017. "Inference for nonprobability samples." *Statistical Science,* 249–264.

Gregoire, Timothy G. 1998. "Design-based and model-based inference in survey sampling: appreciating the difference." *Canadian Journal of Forest Research* 28 (10): 1429–1447.

Groves, Robert M. 2011. "Three eras of survey research." *Public opinion quarterly* 75 (5): 861–871.

Haziza, David, Jean-François Beaumont, et al. 2017. "Construction of weights in surveys: A review." *Statistical Science* 32 (2): 206–226.

Johnson, Timothy P, and Tom W Smith. 2017. "Big data and survey research: Supplement or substitute?" In *Seeing cities through big data,* 113–125. Springer.

Kern, Holger L, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. 2016. "Assessing methods for generalizing experimental impact estimates to target populations." *Journal of research on educational effectiveness* 9 (1): 103–127.

Kim, Jae Kwang, Seho Park, Yilin Chen, and Changbao Wu. 2018. "Combining non-probability and probability survey samples through mass imputation." *arXiv preprint arXiv:1812.10694.*

Lee, Sunghee, and Richard Valliant. 2009. "Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment." *Sociological Methods & Research* 37 (3): 319–343.

Lohr, Sharon L. 2009. *Sampling: design and analysis.* Nelson Education.

Miller, Peter V. 2017. "Is there a future for surveys?" *Public Opinion Quarterly* 81 (S1): 205–212.

Musal, R Muzaffer, Refik Soyer, Christopher McCabe, and Samer A Kharroubi. 2012. "Estimating the population utility function: A parametric Bayesian approach." *European journal of operational research* 218 (2): 538–547.

Nandram, Balgobin. 2007. "Bayesian predictive inference under informative sampling via surrogate samples." *Bayesian statistics and its applications,* 356–374.

Nandram, Balgobin, and Jai Won Choi. 2004. "Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse." *Journal of Nonparametric Statistics* 16 (6): 821–839.

———. 2005. "Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the NHANES data." *Survey Methodology* 31 (1): 73–84.

———. 2010. "A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection." *Journal of the American Statistical Association* 105 (489): 120–135.

Nandram, Balgobin, and Jiani Yin. 2016a. "A nonparametric Bayesian prediction interval for a finite population mean." *Journal of Statistical Computation and Simulation* 86 (16): 3141–3157.

———. 2016b. "Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline." *Statistical Methodology* 28:1–17.

Pfeffermann, Danny, Fernando Antonio Da Silva Moura, and Pedro Luis Do Nascimento Silva. 2006. "Multi-level modelling under informative sampling." *Biometrika* 93 (4): 943–959.

Potthoff, Richard F, Max A Woodbury, and Kenneth G Manton. 1992. ""Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey

weights under superpopulation models." *Journal of the American Statistical Association* 87 (418): 383–396.

Rafei, Ali, Carol AC Flannagan, Brady T West, and Michael R Elliott. 2021. "Robust Bayesian Inference for Big Data: Combining Sensor-based Records with Traditional Survey Data." *arXiv preprint arXiv:2101.07456.*

Rao, JNK. 1966. "Alternative estimators in PPS sampling for multiple characteristics." *Sankhyā: The Indian Journal of Statistics, Series A,* 47–60.

Rao, JNK, et al. 2011. "Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal." *Statistical Science* 26 (2): 240–256.

Rao, JNK. 2020. "On making valid inferences by integrating data from surveys and other sources." *Sankhya B,* 1–31.

Rao, JNK, M Hidiroglou, W Yung, and M Kovacevic. 2010. "Role of weights in descriptive and analytical inferences from survey data: An overview." *Journal of the Indian Society of Agricultural Statistics* 64:129–135.

Royall, Richard M, and Dany Pfeffermann. 1982. "Balanced samples and robust Bayesian inference in finite population sampling." *Biometrika* 69 (2): 401–409.

Särndal, Carl-Erik, Ib Thomsen, Jan M Hoem, DV Lindley, O Barndorff-Nielsen, and Tore Dalenius. 1978. "Design-based and model-based inference in survey sampling [with discussion and reply]." *Scandinavian Journal of Statistics,* 27–52.

Valliant, Richard. 2000. *Finite population sampling and inference: a prediction approach.* 04; QA276. 6, V3.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting elections with non-representative polls." *International Journal of Forecasting* 31 (3): 980–991.

Wendling, T, K Jung, A Callahan, A Schuler, NH Shah, and B Gallego. 2018. "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases." *Statistics in medicine* 37 (23): 3309–3324.