

Article

Bayesian Integration of Probability and Non-Probability Samples

Qi Chen and Balgobin Nandram * 

Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road,
Worcester, MA 01609, USA; qchen6@wpi.edu

* Correspondence: balnan@wpi.edu

Abstract

Non-probability samples (NPS) are increasingly used in practice because they are relatively inexpensive and often contain the study outcome variable. However, NPS may suffer from selection bias and usually do not provide survey weights, making finite-population inference difficult. Probability samples (PS), on the other hand, provide survey weights under known sampling designs but may not contain the study outcome variable. This paper develops a Bayesian approach for integrating probability and non-probability samples under a superpopulation model. The proposed method jointly models latent survey weights and missing probability-sample outcomes within a hierarchical Bayesian framework. The method is compared with a naive Bayesian bootstrap estimator and a weighted regression estimator. A BMI data application and simulation studies show that the proposed method reduces bias relative to the naive estimator while accounting for uncertainty in latent weights and missing outcomes.

Keywords: data integration; Bayesian inference; finite population mean; Gibbs sampler; non-probability samples; probability samples; pseudo-weights; selection bias

1. Introduction

Traditional probability surveys have become increasingly difficult and expensive to conduct because of rising survey costs and declining response rates. As a result, non-probability samples (NPS) are now widely used in practice because they are relatively inexpensive and often contain the study outcome variable. However, NPS generally suffer from selection bias and do not provide survey weights, making finite-population inference challenging. In contrast, probability samples (PS) are collected under known sampling designs and provide survey weights, but the study outcome of interest may not be observed for units in the PS. Individually, neither data source is sufficient for finite-population inference. This setting commonly arises in public health surveys and other large-scale observational studies, where outcome information is often available from non-probability sources, while probability samples remain necessary for valid population inference.

Integrating probability and non-probability samples is difficult because the two data sources contain different types of information. The NPS may contain the study outcome variable, but its uncontrolled selection mechanism can lead to biased inference. The PS, although collected under a valid sampling design, may be too small or may not observe the outcome variable. The main challenge is to combine the two sources while accounting for informative sampling, missing outcomes, and latent survey weights.



Academic Editor: Wei Zhu

Received: 29 May 2026

Revised: 25 June 2026

Accepted: 2 July 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Consider a finite population consisting of N units indexed by $i = 1, \dots, N$, where N may be unknown. Let $\tilde{x}_i$ denote auxiliary covariates and let y_i denote the study outcome for unit i . The inferential goal is to estimate a finite-population quantity such as

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

Suppose a non-probability sample of size n_1 is observed, where $(\tilde{x}_{1i}, y_{1i})$ are available but the survey weights W_{1i} are unknown. Subscript 1 denotes quantities associated with the non-probability sample (NPS). Suppose also that a probability sample of size n_2 is available, where $(\tilde{x}_{2i}, y_{2i})$ are observed but the outcome variable y_{2i} is unavailable. Subscript 2 denotes quantities associated with the probability sample (PS). The goal is to integrate these two incomplete data sources to obtain valid inference for finite-population quantities.

Recent work on integrating probability and non-probability samples has received considerable attention in survey statistics. Elliott and Valliant (2017) [1] reviewed statistical inference for non-probability samples. Beaumont (2020) [2] discussed the increasing role of non-probability data sources in official statistics, and Beaumont and Rao (2021) [3] reviewed methods for combining probability and non-probability samples. Chen, Li, and Wu (2020) [4] proposed a doubly robust approach based on pseudo-weight construction and propensity score adjustment for integrating non-probability samples with probability samples. Rao (2021) [5] proposed using techniques in probability sampling to make valid inference from non-probability samples. Kim et al. (2021) [6] developed a mass-imputation approach for combining probability and non-probability samples. Kim and Morikawa (2023) [7] and Marella (2023) [8] proposed empirical likelihood methods for reducing selection bias in non-probability samples. From a Bayesian perspective, Sakshaug et al. (2019) [9] and Salvatore et al. (2023) [10] developed Bayesian approaches that use information from non-probability samples to improve inference from probability samples. Nandram and Rao (2021, 2023) [11,12] developed Bayesian predictive methods for integrating probability and non-probability samples under informative sampling. Wu (2022) [13] provided a comprehensive review of statistical inference using non-probability samples. Related Bayesian approaches for informative sampling, surrogate sampling, and survey-weight adjustment were also studied in Nandram (2007) [14] and Nandram and Rao (2024) [15].

Many existing approaches rely on a two-stage procedure. In the first stage, survey weights or missing outcomes are estimated. In the second stage, these estimates are treated as known for finite-population inference. For example, Kim et al. (2021) [6] used mass imputation to generate missing information, while Chen, Li, and Wu (2020) [4] constructed pseudo-weights for non-probability samples. Sakshaug et al. (2019) [9], Salvatore et al. (2023) [10], and Nandram and Rao (2021, 2023) [11,12] also follow two-stage Bayesian procedures. Other approaches have used matching methods based on auxiliary covariates and survey weights from the probability sample. Related Bayesian developments incorporating robustness through mixture modeling and flexible outcome–covariate relationships were studied by Nandram (2024) [16] and Lockwood and Nandram (2024) [17], although these methods address different inferential objectives and are beyond the scope of the present paper. As a result, uncertainty associated with the estimated intermediate quantities may not be fully propagated through the final inferential procedure. In contrast, the proposed approach treats the latent survey weights as part of the model and estimates them jointly with the remaining model parameters within a single-stage Bayesian framework. Consequently, uncertainty in the estimated survey weights is directly propagated through the posterior distribution rather than being treated as fixed after a preliminary estimation step.

In this paper, we develop a Bayesian framework for integrating probability and non-probability samples under a superpopulation model. The model incorporates the NPS through a weighted likelihood with latent survey weights and treats the missing PS outcomes as unknown quantities. The proposed method jointly models the latent survey weights and the missing PS outcomes within a single-stage Bayesian framework. As a result, uncertainty in both quantities is directly propagated through the posterior distribution. Posterior computation is carried out using a Gibbs sampling algorithm. We compare the Bayesian approach with a naive Bayesian bootstrap estimator and a weighted regression (REG) approach using both simulation studies and a BMI data application.

The remainder of the paper is organized as follows. Section 2 presents the proposed Bayesian model, prior specification, and alternative estimators. Section 3 develops the posterior computation procedure and Gibbs sampling scheme. Section 4 discusses finite-population mean estimation under the proposed framework. Section 5 presents the BMI data application and model implementation. Section 6 reports the simulation study results. Section 7 provides a discussion of the findings and future research directions.

2. Model Specification

This section introduces the Bayesian model used to integrate the probability and non-probability samples. The non-probability sample weights and the missing probability-sample outcomes are treated as latent quantities. We also describe two comparison methods: a naive Bayesian bootstrap estimator (Naive) and a weighted regression estimator (REG).

2.1. Proposed Bayesian Framework

We assume the outcome variable follows a linear superpopulation regression model,

$$y_i | x_i, \beta, \sigma^2 \sim \mathcal{N}(x_i' \beta, \sigma^2), \quad i = 1, \dots, N.$$

Here, β denotes the vector of regression coefficients and σ^2 denotes the common variance. The superpopulation model links the probability and non-probability samples through a common regression structure. It also allows prediction of the missing probability-sample outcomes.

To account for heterogeneity and potential selection effects in the non-probability sample, we model its contribution using a weighted likelihood based on latent survey weights. Let W_{1i} denote the original unobserved survey weight for the i th unit in the non-probability sample. Following Nandram and Rao (2021, 2023) [11,12], we define the adjusted survey weight

$$w_{1i} = \hat{n}_1 \frac{W_{1i}}{\sum_{j=1}^{n_1} W_{1j}}, \quad i = 1, \dots, n_1,$$

where $\hat{n}_1$ is the effective sample size of the non-probability sample, given by

$$\hat{n}_1 = \frac{(\sum_{i=1}^{n_1} W_{1i})^2}{\sum_{i=1}^{n_1} W_{1i}^2},$$

which is the standard effective sample size definition for weighted survey samples introduced by Kish (1965) [18]. When the survey weights are homogeneous, the coefficient of variation of the survey weights is zero, and the effective sample size equals the actual sample size. As the variability of the weights increases, the effective sample size decreases, reflecting the loss of information due to unequal weighting. Consequently, extreme survey weights lead to a smaller effective sample size and greater posterior uncertainty.

In practice, the original survey weights W_{1i} for the non-probability sample are not observed. However, pseudo weights can be constructed using existing approaches such as matching non-probability sample units to probability sample units based on auxiliary covariates or calibration-based methods proposed by Chen, Li, and Wu (2020) [4]. These pseudo weights are used only to compute the effective sample size $\hat{n}_1$. The effective sample size $\hat{n}_1$ is computed prior to posterior sampling and treated as fixed during posterior computation. Because the proposed framework involves several types of survey weights and sample-size quantities, a notation summary is provided in Appendix A.

Given the adjusted weights w_{1i} , the likelihood contribution of the non-probability sample is

$$L_1(\beta, \sigma^2 \mid y_1, x_1, w_1) \propto \prod_{i=1}^{n_1} \left[\mathcal{N}(y_{1i} \mid x_{1i}'\beta, \sigma^2) \right]^{w_{1i}}.$$

Equivalently,

$$y_{1i} \mid x_{1i}, \beta, \sigma^2, w_{1i} \sim \mathcal{N}\left(x_{1i}'\beta, \frac{\sigma^2}{w_{1i}}\right).$$

The probability sample is drawn under a known sampling design, and its likelihood contribution is weighted by the observed survey weights. Let W_{2i} denote the original survey weight for the i th unit in the probability sample. We define the adjusted survey weight

$$w_{2i} = \hat{n}_2 \frac{W_{2i}}{\sum_{j=1}^{n_2} W_{2j}}, \quad i = 1, \dots, n_2,$$

where $\hat{n}_2$ is the effective sample size of the probability sample, given by

$$\hat{n}_2 = \frac{(\sum_{i=1}^{n_2} W_{2i})^2}{\sum_{i=1}^{n_2} W_{2i}^2}.$$

Since the original survey weights W_{2i} are observed for the probability sample, the adjusted weights w_{2i} are treated as fixed.

The outcome y_{2i} is unobserved for units in the probability sample and is treated as latent data. Under the superpopulation model, the likelihood contribution of the probability sample is

$$L_2(\beta, \sigma^2 \mid y_2, x_2, w_2) \propto \prod_{i=1}^{n_2} \left[\mathcal{N}(y_{2i} \mid x_{2i}'\beta, \sigma^2) \right]^{w_{2i}}.$$

Equivalently,

$$y_{2i} \mid x_{2i}, \beta, \sigma^2, w_{2i} \sim \mathcal{N}\left(x_{2i}'\beta, \frac{\sigma^2}{w_{2i}}\right).$$

2.2. Prior Specification

We assign the non-informative prior

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

For the adjusted non-probability sample weights, we assume

$$w_{1i} \sim \text{Gamma}(\alpha_w, \beta_{wi}), \quad i = 1, \dots, n_1,$$

subject to the calibration constraint

$$\sum_{i=1}^{n_1} w_{1i} = \hat{n}_1.$$

Since the probability and non-probability samples are drawn from the same finite population, we use the observed probability sample weights to estimate the distribution of the non-probability sample weights. Rather than specifying a Gamma model directly, we first model the normalized adjusted weights on the log scale. This transformation preserves positivity, reduces the skewness commonly observed in survey weights, and allows the relationship between the auxiliary covariates and the weights to be described through a standard normal linear model. The resulting lognormal distribution is then approximated by a Gamma distribution using moment matching. This approximation preserves positivity while yielding a convenient form for posterior computation. In particular, it leads to closed form full conditional distributions for the adjusted weights within the Gibbs sampler, thereby simplifying computation and improving computational efficiency. Specifically, we assume

$$\log\left(\frac{w_{2i}}{\hat{n}_2}\right) \sim \mathcal{N}(x_{2i}'\gamma_0, \tau_0^2).$$

and obtain estimates $\hat{\gamma}_0$ and $\hat{\tau}_0^2$ by fitting this model to the adjusted probability-sample weights.

We then assume that a similar relationship holds for the non-probability sample:

$$\log\left(\frac{w_{1i}}{\hat{n}_1}\right) \sim \mathcal{N}(x_{1i}'\hat{\gamma}_0, \hat{\tau}_0^2).$$

Normalizing the adjusted weights by the corresponding effective sample sizes, $\hat{n}_1$ and $\hat{n}_2$, allows the model to focus on relative weight patterns rather than absolute weight magnitudes. Consequently, the proposed framework does not assume that the raw weights in the probability and non-probability samples are identical. Instead, it assumes that the relationship between the auxiliary covariates and the normalized adjusted weights can be described through a common covariate structure.

This implies that w_{1i} follows a lognormal distribution with mean and variance determined by x_{1i} .

The resulting lognormal distribution is then approximated by a Gamma distribution using moment matching. This approximation preserves positivity while yielding a closed form for posterior computation. In particular,

$$\alpha_w = \frac{1}{\exp(\hat{\tau}_0^2) - 1},$$

while the rate parameter varies with the covariates,

$$\beta_{wi} = \frac{\exp(-x_{1i}'\hat{\gamma}_0 - \log(\hat{n}_1) - \hat{\tau}_0^2/2)}{\exp(\hat{\tau}_0^2) - 1}.$$

For the probability sample, where y_{2i} is not observed, we treat the missing outcomes as unknown quantities. Following the superpopulation model, we use the prior predictive distribution

$$y_{2i} \mid x_{2i}, \beta_0, \sigma_0^2 \sim \mathcal{N}(x_{2i}'\beta_0, \sigma_0^2), \quad i = 1, \dots, n_2,$$

where (β_0, σ_0^2) are hyperparameters obtained by fitting the model

$$y_{1i} \sim \mathcal{N}(x_{1i}'\beta_0, \sigma_0^2), \quad i = 1, \dots, n_1,$$

using the observed non-probability sample data.

The prior predictive distribution provides information about the missing probability sample outcomes through the superpopulation model estimated from the observed non-probability sample. However, the missing outcomes are not directly replaced by observed non-probability sample outcomes. Instead, inference for the latent outcomes combines information from the superpopulation model, the probability sample covariates and survey weights through the joint posterior distribution. Consequently, potential bias in the non-probability sample is not automatically transferred to the probability sample, because posterior inference is informed by both data sources rather than solely by the non-probability sample.

Let $\tilde{y}_2 = (y_{2i}, i = 1, \dots, n_2)$ denote the unobserved outcomes in the probability sample. Given the observed data $(\tilde{y}_1, \tilde{x}_1, \tilde{x}_2, \tilde{w}_2)$, the joint posterior density of $(\tilde{\beta}, \tilde{\sigma}^2, \tilde{w}_1, \tilde{y}_2)$ is proportional to

$$\begin{aligned} \pi(\tilde{\beta}, \tilde{\sigma}^2, \tilde{w}_1, \tilde{y}_2 \mid \tilde{y}_1, \tilde{x}_1, \tilde{x}_2, \tilde{w}_2) & \propto \underbrace{\prod_{i=1}^{n_1} \left[\frac{1}{\sqrt{2\pi(\sigma^2/w_{1i})}} \exp \left\{ -\frac{(y_{1i} - x_{1i}'\tilde{\beta})^2}{2(\sigma^2/w_{1i})} \right\} \right]}_{\text{NPS likelihood contribution}} \times \underbrace{\prod_{i=1}^{n_2} \left[\frac{1}{\sqrt{2\pi(\sigma^2/w_{2i})}} \exp \left\{ -\frac{(y_{2i} - x_{2i}'\tilde{\beta})^2}{2(\sigma^2/w_{2i})} \right\} \right]}_{\text{PS likelihood contribution}} \\ & \times \underbrace{\frac{1}{\sigma^2}}_{\text{prior for } (\beta, \sigma^2)} \times \underbrace{\prod_{i=1}^{n_1} \frac{\beta_{wi}^{\alpha_w}}{\Gamma(\alpha_w)} w_{1i}^{\alpha_w-1} \exp\{-\beta_{wi} w_{1i}\}}_{\text{prior for } w_1} \times \underbrace{\mathbb{I}\left(\sum_{i=1}^{n_1} w_{1i} = \hat{n}_1\right)}_{\text{calibration constraint}} \\ & \times \underbrace{\prod_{i=1}^{n_2} \left[\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left\{ -\frac{(y_{2i} - x_{2i}'\tilde{\beta}_0)^2}{2\sigma_0^2} \right\} \right]}_{\text{prior predictive for } \tilde{y}_2}. \end{aligned}$$

Equivalently, collecting terms, the posterior density can be written as

$$\begin{aligned} \pi(\tilde{\beta}, \tilde{\sigma}^2, \tilde{w}_1, \tilde{y}_2 \mid \tilde{y}_1, \tilde{x}_1, \tilde{x}_2, \tilde{w}_2) & \propto (\sigma^2)^{-\frac{(n_1+n_2)}{2}-1} \prod_{i=1}^{n_1} w_{1i}^{\alpha_w-\frac{1}{2}} \exp \left\{ -\sum_{i=1}^{n_1} \beta_{wi} w_{1i} \right\} \mathbb{I}\left(\sum_{i=1}^{n_1} w_{1i} = \hat{n}_1\right) \\ & \times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_1} w_{1i} (y_{1i} - x_{1i}'\tilde{\beta})^2 \right\} \times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_2} w_{2i} (y_{2i} - x_{2i}'\tilde{\beta})^2 \right\} \\ & \times \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n_2} (y_{2i} - x_{2i}'\tilde{\beta}_0)^2 \right\}. \end{aligned}$$

The posterior density combines information from the observed non-probability sample outcomes, the latent probability-sample outcomes, and the prior distributions for the latent weights and missing outcomes.

2.3. Alternative Estimators

As benchmarks, we consider two alternative estimators. The first is a naive Bayesian bootstrap estimator based on directly combining the observed outcomes from the non-probability sample and the probability sample. In the simulation study, the outcome

variable is available for units in both samples before the probability-sample outcomes are treated as missing. Let

$$\underset{\sim}{y} = (y_{11}, \dots, y_{1n_1}, y_{21}, \dots, y_{2n_2})$$

denote the combined observed outcomes. At each bootstrap iteration, random weights

$$(p_1, \dots, p_n) \sim \text{Dirichlet}(1, \dots, 1)$$

are generated, and the population mean is estimated by

$$\bar{Y}_{\text{naive}} = \sum_{i=1}^n p_i y_i.$$

This estimator does not account for the sampling mechanism or survey weights. Posterior summaries and 95% HPD intervals are computed from the resulting bootstrap draws.

The second comparison method is a weighted regression estimator based only on the non-probability sample. This method accounts for heterogeneity in the non-probability sample through adjusted weights, but it does not explicitly model the probability sample. For the non-probability sample, we assume

$$y_{1i} \mid \underset{\sim}{\beta}, \sigma^2 \sim \mathcal{N}\left(\underset{\sim}{x}_{1i}' \underset{\sim}{\beta}, \frac{\sigma^2}{w_{1i}}\right), \quad i = 1, \dots, n_1,$$

where the adjusted weights w_{1i} are obtained using the same procedure as in the proposed Bayesian model, based on the probability-sample survey weights W_{2i} .

We assign the prior

$$\pi(\underset{\sim}{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Under this model, the posterior distribution is proportional to

$$\pi(\underset{\sim}{\beta}, \sigma^2 \mid \underset{\sim}{y}_1, \underset{\sim}{w}_1) \propto (\sigma^2)^{-\frac{n_1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n_1} w_{1i} \left(y_{1i} - \underset{\sim}{x}_{1i}' \underset{\sim}{\beta}\right)^2\right\}.$$

Define

$$A_{\beta_{\text{REG}}} = \sum_{i=1}^{n_1} w_{1i} \underset{\sim}{x}_{1i} \underset{\sim}{x}_{1i}', \quad B_{\beta_{\text{REG}}} = \sum_{i=1}^{n_1} w_{1i} \underset{\sim}{x}_{1i} y_{1i}.$$

Then the posterior distributions have closed-form expressions:

$$\underset{\sim}{\beta} \mid \sigma^2, \underset{\sim}{y}_1, \underset{\sim}{w}_1 \sim \mathcal{N}(\mu_{\beta_{\text{REG}}}, \Sigma_{\beta_{\text{REG}}}),$$

with

$$\mu_{\beta_{\text{REG}}} = A_{\beta_{\text{REG}}}^{-1} B_{\beta_{\text{REG}}}, \quad \Sigma_{\beta_{\text{REG}}} = \sigma^2 A_{\beta_{\text{REG}}}^{-1},$$

and

$$\sigma^2 \mid \underset{\sim}{y}_1, \underset{\sim}{w}_1 \sim \text{IG}\left(\frac{n_1 - p}{2}, \frac{Q(\mu_{\beta_{\text{REG}}})}{2}\right),$$

where

$$Q(\mu_{\beta_{\text{REG}}}) = \sum_{i=1}^{n_1} w_{1i} \left(y_{1i} - \underset{\sim}{x}_{1i}' \mu_{\beta_{\text{REG}}}\right)^2,$$

and p is the dimension of $\underset{\sim}{\beta}$.

Posterior samples can therefore be obtained by first drawing σ^2 from its marginal posterior distribution and then drawing $\underset{\sim}{\beta}$ conditional on σ^2 . Unlike the proposed Bayesian

model, the REG approach treats the adjusted weights as fixed and uses only the non-probability sample outcomes for inference.

3. Posterior Inference and Computation

The REG model has closed-form posterior distributions and does not require MCMC computation. Therefore, we focus on posterior computation for the proposed Bayesian model. In this section, we derive the full conditional distributions for all unknown quantities and describe the Gibbs sampling algorithm. The full conditional distributions used in the Gibbs sampler are summarized below:

- $\beta | \cdot$: multivariate normal,
- $\sigma^2 | \cdot$: inverse-gamma,
- $y_{2i} | \cdot$: normal,
- $w_{1i} | \cdot$: gamma (followed by normalization to satisfy calibration).

3.1. Full Conditional Posterior Distributions

For β , conditioning on the rest, the kernel is

$$\pi(\beta | \cdot) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n_1} w_{1i} (y_{1i} - x_{1i}' \beta)^2 + \sum_{i=1}^{n_2} w_{2i} (y_{2i} - x_{2i}' \beta)^2 \right] \right\}.$$

Define

$$A_\beta = \sum_{i=1}^{n_1} w_{1i} x_{1i} x_{1i}' + \sum_{i=1}^{n_2} w_{2i} x_{2i} x_{2i}', \quad B_\beta = \sum_{i=1}^{n_1} w_{1i} x_{1i} y_{1i} + \sum_{i=1}^{n_2} w_{2i} x_{2i} y_{2i}.$$

Then the full conditional distribution is multivariate normal,

$$\beta | \sigma^2, w_1, y_2, y_1, x_1, x_2, w_2 \sim \mathcal{N}(\mu_\beta, \Sigma_\beta),$$

with

$$\Sigma_\beta = \sigma^2 A_\beta^{-1}, \quad \mu_\beta = A_\beta^{-1} B_\beta.$$

For σ^2 , conditioning on the rest, the kernel is

$$\pi(\sigma^2 | \cdot) \propto (\sigma^2)^{-\frac{n_1+n_2}{2}-1} \exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n_1} w_{1i} (y_{1i} - x_{1i}' \beta)^2 + \sum_{i=1}^{n_2} w_{2i} (y_{2i} - x_{2i}' \beta)^2 \right] \right\}$$

Let

$$Q(\beta, y_2 | w_1, y_1) = \sum_{i=1}^{n_1} w_{1i} (y_{1i} - x_{1i}' \beta)^2 + \sum_{i=1}^{n_2} w_{2i} (y_{2i} - x_{2i}' \beta)^2.$$

Then the full conditional distribution is inverse-gamma,

$$\sigma^2 | \beta, w_1, y_2, y_1, x_1, x_2, w_2 \sim \text{IG} \left(\frac{n_1 + n_2}{2}, \frac{Q(\beta, y_2 | w_1, y_1)}{2} \right).$$

For each $i = 1, \dots, n_2$, the kernel for a single missing probability-sample outcome y_{2i} is

$$\pi(y_{2i} | \cdot) \propto \exp \left\{ -\frac{w_{2i}}{2\sigma^2} (y_{2i} - x_{2i}' \beta)^2 \right\} \exp \left\{ -\frac{1}{2\sigma_0^2} (y_{2i} - x_{2i}' \beta_0)^2 \right\}$$

This is the product of two Gaussian kernels, and therefore the result is another normal distribution. The full conditional distribution is

$$y_{2i} \mid \beta, \sigma^2, w_{2i}, x_{2i}, \beta_0, \sigma_0^2 \sim \mathcal{N}(\mu_{2i}, v_{2i}),$$

where

$$v_{2i} = \left(\frac{w_{2i}}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1}, \quad \mu_{2i} = v_{2i} \left(\frac{w_{2i} x_{2i}' \beta}{\sigma^2} + \frac{x_{2i}' \beta_0}{\sigma_0^2} \right).$$

For each $i = 1, \dots, n_1$, the kernel for a single adjusted NPS weight w_{1i} is

$$\pi(w_{1i} \mid \cdot) \propto \left(w_{1i}^{\alpha_w - \frac{1}{2}} \right) \times \exp \left\{ -w_{1i} \left[\beta_{wi} + \frac{1}{2\sigma^2} (y_{1i} - x_{1i}' \beta)^2 \right] \right\} \times \mathbb{I} \left(\sum_{i=1}^{n_1} w_{1i} = \hat{n}_1 \right)$$

Without the constraint $\mathbb{I}(\sum_{i=1}^{n_1} w_{1i} = \hat{n}_1)$, the full conditional kernel in w_{1i} is gamma,

$$w_{1i} \mid \beta, \sigma^2, y_{1i}, x_{1i} \sim \text{Gamma} \left(\alpha_w + \frac{1}{2}, \beta_{wi} + \frac{(y_{1i} - x_{1i}' \beta)^2}{2\sigma^2} \right).$$

In the Gibbs sampler, after drawing w_{1i} for $i = 1, \dots, n_1$ from the gamma distribution above, the weights are rescaled to satisfy the calibration constraint. This normalization preserves positivity of the sampled weights and ensures consistency with the effective sample size used in the calibration constraint.

3.2. Gibbs Sampling Scheme

Posterior inference for the proposed Bayesian model is carried out using a Gibbs sampling scheme. We summarize the Gibbs sampling steps below.

At iteration $t = 1, \dots, T$, given current values $(\beta^{(t-1)}, (\sigma^2)^{(t-1)}, y_2^{(t-1)}, w_1^{(t-1)})$:

1. Although the previous subsection presents the separate full conditional posterior distributions for β and σ^2 , in implementation we update (β, σ^2) jointly using a block Gibbs step. Using the factorization

$$p(\beta, \sigma^2 \mid \cdot) = p(\beta \mid \sigma^2, \cdot) p(\sigma^2 \mid \cdot),$$

the block update can be implemented sequentially as follows:

- (a) Sample $(\sigma^2)^{(t)}$ from its marginal conditional distribution, obtained by integrating out β from the joint conditional distribution of (β, σ^2) .
 - (b) Sample $\beta^{(t)}$ from the normal full conditional distribution using $(\sigma^2)^{(t)}$.
2. For $i = 1, \dots, n_2$, sample $y_{2i}^{(t)}$ from the normal full conditional distribution.
 3. For $i = 1, \dots, n_1$, sample $w_{1i}^{(t)}$ from the gamma full conditional distribution, then rescale $(w_{11}^{(t)}, \dots, w_{1n_1}^{(t)})$ to satisfy $\sum_{i=1}^{n_1} w_{1i}^{(t)} = \hat{n}_1$. That is, each sampled weight $w_{1i}^{(t)}$ is replaced by

$$w_{1i}^{(t)} = \hat{n}_1 \frac{w_{1i}^{(t)}}{\sum_{j=1}^{n_1} w_{1j}^{(t)}}, \quad i = 1, \dots, n_1.$$

Posterior summaries are obtained from draws after burn-in and thinning.

4. Population Mean Estimation

This section describes estimation of the finite-population mean under the proposed Bayesian model and the REG model. Under the superpopulation model, the finite-population mean has a conditional normal distribution given the model parameters. Because the population covariates are not fully observed, the probability sample and its survey weights are used to estimate the population mean of the covariates. Posterior inference for the finite-population mean is obtained from posterior draws of the model parameters.

4.1. Model-Based Population Prediction

Under the superpopulation model specified in Section 2.1, the population outcomes satisfy

$$y_i \mid x_i, \beta, \sigma^2 \sim \mathcal{N}(x_i' \beta, \sigma^2), \quad i = 1, \dots, N,$$

independently across units. Given (β, σ^2) , the finite population $\{y_1, \dots, y_N\}$ consists of independent normal random variables.

The population total satisfies

$$\sum_{i=1}^N y_i \sim \mathcal{N}\left(\sum_{i=1}^N x_i' \beta, N\sigma^2\right),$$

and the population mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$$

has conditional distribution

$$\bar{Y} \mid \beta, \sigma^2 \sim \mathcal{N}\left(\bar{X}' \beta, \frac{\sigma^2}{N}\right), \quad \bar{X} = \frac{1}{N} \sum_{i=1}^N x_i.$$

4.2. Estimation and Uncertainty Quantification

In practice, covariates x_i are not observed for all population units. We therefore estimate the population mean of the covariates using the probability sample. Specifically, we approximate

$$\bar{x} = \frac{\sum_{i=1}^{n_2} W_{2i} x_{2i}}{\sum_{i=1}^{n_2} W_{2i}}, \quad N = \sum_{i=1}^{n_2} W_{2i},$$

where W_{2i} are the known survey weights from the probability sample. Substituting these quantities into the model-based distribution gives

$$\bar{Y} \mid \beta, \sigma^2 \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n_2} W_{2i} x_{2i}}{\sum_{i=1}^{n_2} W_{2i}} \beta, \frac{\sigma^2}{\sum_{i=1}^{n_2} W_{2i}}\right).$$

This formulation is used for both the proposed Bayesian model and the REG model. The difference lies in how the parameters (β, σ^2) are obtained under each model. For the proposed Bayesian model, posterior samples of (β, σ^2) are obtained using Gibbs sampling. For the REG model, posterior samples are obtained from the corresponding closed-form posterior distributions based on the NPS. The naive estimator is based on a Bayesian bootstrap using the combined observed outcomes from the NPS and PS samples and does not use the model-based formulation above.

Posterior samples of $\tilde{Y}$ are obtained by drawing (β, σ^2) from their joint posterior distribution and then sampling $\tilde{Y}$ from the conditional normal distribution. Posterior summaries include point estimates and 95% HPD intervals.

5. Application to BMI Data

This section illustrates the proposed Bayesian framework using a BMI dataset. The application considers a setting where outcome information is available from a non-probability sample, while survey weights are available from a probability sample. We describe the BMI dataset, construct the probability and non-probability samples, and implement the naive, REG, and proposed Bayesian methods. The resulting finite-population mean estimates are then compared.

5.1. Data Description

We illustrate the proposed method using a BMI dataset consisting of 1867 sampled units. The dataset includes the variables County, NPS_PS, Obs, Age, Race, Sex, BMI, and Weights. BMI is the outcome variable of interest. Age, Race, and Sex are used as auxiliary covariates in the superpopulation model. Weights contains survey weights, and NPS_PS indicates whether an observation belongs to the non-probability sample or the probability sample. Obs indexes observations within each county, and County identifies the county for each sampled unit.

In the original dataset, both BMI and survey weights are observed for all sampled units. To mimic the data integration setting considered in this paper, we partition the data into a non-probability sample and a probability sample. For the non-probability sample, the survey weights are removed, while for the probability sample, BMI is treated as missing. This setup represents situations where outcome information is available from non-probability samples, while probability samples provide survey weights.

The main inferential goal is to estimate the finite-population mean of BMI. Because the population mean can be sensitive to extreme BMI values, we also report the weighted population median computed from the full dataset as a reference summary measure. The median is less affected by unusually large or small BMI values and is therefore useful when the outcome distribution exhibits some degree of skewness.

Figure 1 presents graphical summaries of the sampled BMI data. The distribution exhibits moderate right skewness and several large observations, indicating that the normality assumption is only an approximation. Nevertheless, the skewness is not severe, and the normal model provides a reasonable working model for the proposed Bayesian framework.

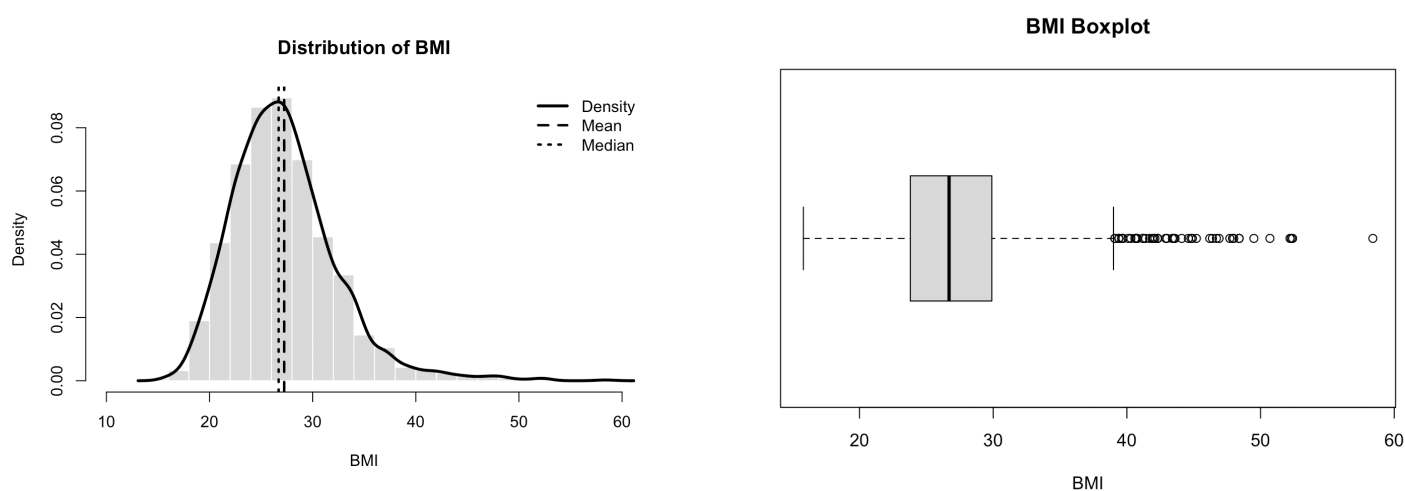


Figure 1. Distribution of BMI in the sampled data. (Left) histogram. (Right) boxplot.

5.2. Model Implementation

We apply the three methods described in Section 2 to the BMI dataset: the naive estimator, the REG model, and the proposed Bayesian model.

For the naive method, we apply a Bayesian bootstrap procedure to the combined observed outcomes from the non-probability sample and the probability sample. Posterior summaries and HPD intervals are obtained from the bootstrap draws. This is feasible in the current application since BMI is observed for all units in the original dataset. This method ignores the sampling mechanism and survey weights and is used as a benchmark estimator.

For the REG model, we fit the weighted regression model using the non-probability sample and obtain closed-form posterior distributions for (β, σ^2) . These parameter estimates are then used in the model-based framework described in Section 4 to obtain inference for the population mean.

For the proposed Bayesian model, posterior inference is carried out using the Gibbs sampling scheme described in Section 3. Specifically, we run the Gibbs sampler for 70,000 iterations, discarding the first 34,000 iterations as burn-in and retaining every 36th draw for posterior inference. This results in 1000 retained posterior samples. Posterior summaries are computed based on the retained samples.

To assess the performance of the Gibbs sampler, we examined trace plots, Geweke diagnostics, and effective sample sizes (ESS) for the main model parameters and the finite-population mean.

Figure 2 presents trace plots for the regression coefficients, variance parameter, and finite-population mean. The chains mix well without noticeable trends or nonstationary behavior, suggesting satisfactory convergence of the Gibbs sampler.

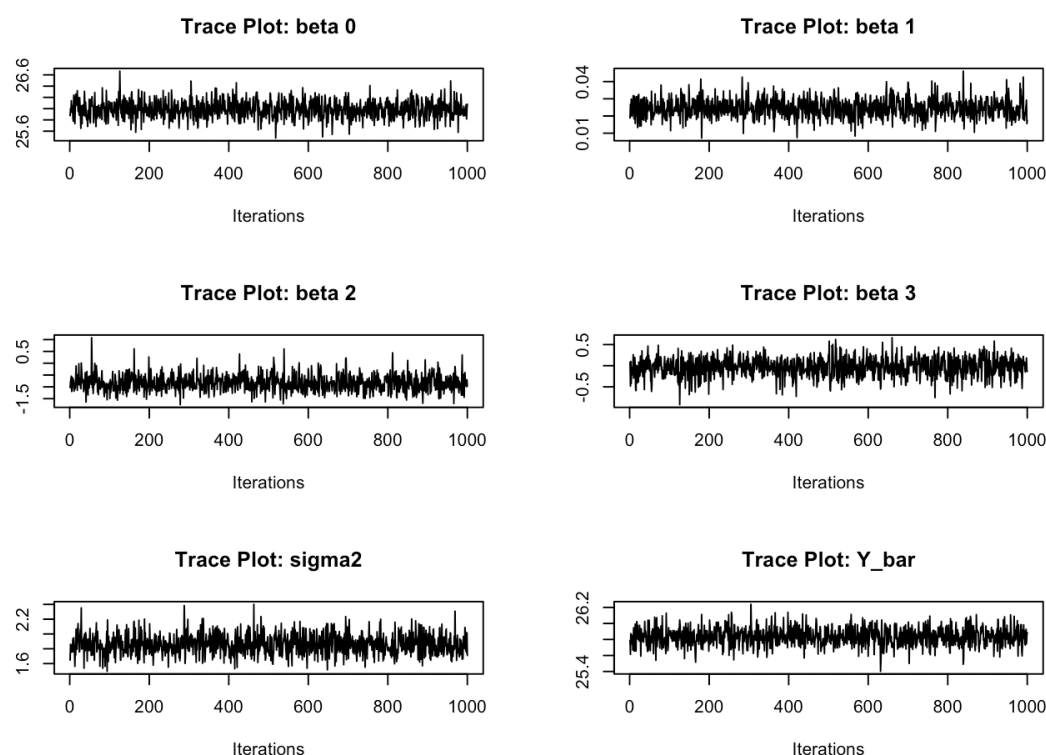


Figure 2. Trace plots for selected parameters and $\bar{Y}$ under the proposed Bayesian model.

Table 1 summarizes the ESS values and Geweke diagnostic p -values for selected parameters. All parameters have effective sample sizes of at least 1000, and all Geweke p -values are greater than 0.05, indicating satisfactory convergence of the Gibbs sampler. Combined

with the trace plots in Figure 2, these diagnostics suggest satisfactory convergence and mixing of the Gibbs sampler.

Table 1. MCMC diagnostics for the proposed Bayesian model.

Parameter	ESS	Geweke p -Value
β_0	1381	0.8259
β_1	1175	0.2678
β_2	1031	0.4042
β_3	1000	0.4923
σ^2	1000	0.4164
$\bar{Y}$	1111	0.8745

5.3. Estimation of Population Mean

For the REG and proposed Bayesian methods, inference for the finite-population mean is obtained using the framework described in Section 4. The naive method uses a Bayesian bootstrap estimator based on the combined observed outcomes from the NPS and PS samples.

Table 2 summarizes the estimated population mean of BMI under the naive, REG, and proposed Bayesian methods. In addition, the weighted mean and weighted median computed from the full dataset are included as reference values. The weighted median is included because the population mean may be sensitive to extreme BMI values. All interval estimates reported in Table 2 are 95% HPD intervals.

The naive estimator overestimates the population mean because it directly combines the non-probability and probability samples without accounting for the sampling mechanism. The REG estimator partially adjusts for this bias and produces an estimate closer to the reference mean. The proposed Bayesian method produces a lower estimate than both the naive and REG estimators because it explicitly accounts for uncertainty in the latent survey weights and the unobserved PS outcomes through posterior inference. By modeling the selection mechanism and propagating uncertainty throughout the analysis, the proposed method provides a more comprehensive adjustment for selection bias.

The reference value is computed from the sampled dataset rather than the true finite population. Therefore, some discrepancy between the interval estimates and the reference value is expected. For the naive and REG methods, both the weighted mean (26.6721) and the weighted median (26.1000) lie to the left of the 95% HPD intervals. For the proposed Bayesian method, both reference values lie to the right of the 95% HPD interval. This result indicates that the proposed method provides a stronger adjustment for selection bias than the naive and REG methods.

Table 2. Comparison of estimated population means under different methods.

Method	Reference Mean	Reference Median	Estimate	Lower 95%	Upper 95%
Naive	26.6721	26.1000	27.2420	27.0175	27.4647
REG	26.6721	26.1000	27.0933	26.8128	27.3268
Proposed	26.6721	26.1000	25.8367	25.6094	26.0573

6. Simulation Study

We conduct a simulation study to evaluate estimation of the finite-population mean under the proposed Bayesian method. The proposed method is compared with two benchmark estimators: the naive estimator and the REG estimator.

6.1. Simulation Design

The simulation design follows Nandram and Rao (2021) [11], who adapted the design-based simulation framework of Chen, Li, and Wu (2020) [4] to a BMI data setting.

We first generate a finite population of size $N = 10,000$. For each population unit $i = 1, \dots, N$, three covariates are generated. The first covariate is generated from

$$x_{1i} \sim \text{Uniform}(20, 90).$$

The second covariate is generated from a Bernoulli distribution with probability

$$P(x_{2i} = 1 \mid x_{1i}) = \frac{\exp(b_{2i})}{1 + \exp(b_{2i})}, \quad b_{2i} = (23.8449 + 0.0559x_{1i})^{1/10}.$$

The third covariate is generated from a Bernoulli distribution with probability

$$P(x_{3i} = 1 \mid x_{1i}, x_{2i}) = \frac{\exp(b_{3i})}{1 + \exp(b_{3i})}, \quad \text{where } b_{3i} = (23.8449 + 0.0559x_{1i} + 2.2656x_{2i})^{1/10}.$$

Given the covariates, the outcome variable is generated from the mixture model

$$y_i \mid x_{1i}, x_{2i}, x_{3i} \sim (1 - p)N(\eta_i, \gamma\sigma^2) + pN(\eta_i, \sigma^2),$$

where,

$$\eta_i = 23.8449 + 0.0559x_{1i} + 2.2656x_{2i} + 0.2525x_{3i}.$$

We choose $p = 0.1$ and $\gamma = 0.5$, so that 90% of the observations are generated from $N(\eta_i, 0.5\sigma^2)$ and 10% are generated from $N(\eta_i, \sigma^2)$.

The value of σ is chosen so that the correlation between η_i and y_i is approximately equal to a pre-specified value ρ . Details of the derivation are given in Appendix B. We consider

$$\rho \in \{0.2, 0.3, 0.5, 0.7, 0.8\}.$$

The finite-population mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$$

is treated as the true value for evaluating the estimators.

For each generated finite population, we draw a non-probability sample and a probability sample. The non-probability sample is generated using Poisson sampling with inclusion probability

$$\pi_{1i} = \frac{\exp(\theta_0 + 0.1x_{1i} + 0.2x_{2i} + 0.1x_{3i})}{1 + \exp(\theta_0 + 0.1x_{1i} + 0.2x_{2i} + 0.1x_{3i})}.$$

We consider two sample size scenarios,

$$(n_1, n_2) = (800, 200)$$

and

$$(n_1, n_2) = (400, 100).$$

For each value of ρ , the same generated finite population is used for both scenarios, so that differences in performance can be attributed solely to the sample sizes. For each scenario, the intercept θ_0 is chosen so that the expected NPS sample size is approximately 800 and 400, respectively. Thus, the realized NPS sample size varies across replications.

The probability sample is selected using randomized systematic probability proportional to size sampling. The inclusion probability is

$$\pi_{2i} = \frac{n_2 z_i}{\sum_{j=1}^N z_j},$$

where n_2 denotes the probability sample size and takes values 200 and 100 under the two scenarios, respectively, and

$$z_i = \theta_1 + x_{1i} + 0.2x_{2i} + 0.1x_{3i}.$$

The value of θ_1 is chosen so that the ratio of the largest to the smallest size measure is approximately 50. The corresponding probability sample survey weight is

$$W_{2i} = \frac{1}{\pi_{2i}}.$$

In the observed data used for estimation, the NPS contains $(x_{1i}, x_{2i}, x_{3i}, y_i)$, but the survey weights are treated as unavailable. The PS contains $(x_{1i}, x_{2i}, x_{3i}, W_{2i})$, but the outcome y_i is treated as missing.

6.2. Comparison of the Models

For each sample size scenario, we compare three methods: the naive estimator, the REG estimator, and the proposed Bayesian method. The naive method applies a Bayesian bootstrap procedure to the combined NPS and PS outcomes.

The REG method uses the NPS outcome data together with adjusted pseudo-weights constructed from the probability sample. The proposed Bayesian method treats the adjusted NPS weights and the missing PS outcomes as unknown quantities and updates them within the Gibbs sampler.

We report several posterior summary measures for model comparison. The posterior mean (PM) is the posterior average of the finite-population mean estimator. The posterior standard deviation (PSD) measures posterior uncertainty. The coefficient of variation (CV) is defined as the ratio of the posterior standard deviation to the posterior mean. The numerical standard error (NSE) measures Monte Carlo variability due to posterior simulation.

Let $\hat{Y}^{(b)}$ denote the estimate of the finite-population mean in replication b , and let $\bar{Y}$ denote the true finite-population mean. Across replications, we compute the absolute relative bias (ARB),

$$\text{ARB} = \frac{1}{B} \sum_{b=1}^B \left| \frac{\hat{Y}^{(b)} - \bar{Y}}{\bar{Y}} \right|,$$

the posterior root mean squared error (PRMSE),

$$\text{PRMSE} = \frac{1}{B} \sum_{b=1}^B \sqrt{\left(\hat{Y}^{(b)} - \bar{Y} \right)^2 + \widehat{\text{Var}}(\bar{Y} \mid \text{data}^{(b)})},$$

the empirical coverage probability (Cov),

$$\text{Cov} = \frac{1}{B} \sum_{b=1}^B I\left(L^{(b)} \leq \bar{Y} \leq U^{(b)}\right),$$

and the average interval width (Wid),

$$\text{Wid} = \frac{1}{B} \sum_{b=1}^B \left(U^{(b)} - L^{(b)} \right).$$

Smaller values of ARB and PRMSE correspond to improved point estimation accuracy. Higher coverage probabilities and smaller interval widths indicate improved interval estimation performance.

6.3. Simulation Results

Table 3 reports the target and realized values of ρ , the corresponding value of σ , and the true finite-population mean. The realized correlations are close to the target values, showing that the simulation design adequately controls the association between the covariates and the outcome. As ρ increases, the value of σ decreases, reflecting a stronger relationship between the outcome and the covariates.

Table 3. Diagnostics for the generated finite populations under different values of ρ .

Target ρ	Realized ρ	σ	True Mean
0.2	0.1945	9.5642	28.8460
0.3	0.3175	6.2079	28.9188
0.5	0.4893	3.3815	28.9088
0.7	0.7002	1.9917	28.9196
0.8	0.7988	1.4642	28.9164

Table 4 reports posterior summaries for the finite-population mean under the two sample size scenarios. The naive estimator consistently overestimates the true finite-population mean across all values of ρ . The naive estimator is included as a benchmark method and does not adjust for the sampling mechanism, survey weights, or selection bias. Consequently, units with larger outcomes tend to be overrepresented in the sample, resulting in upwardly biased estimates of the population mean. In contrast, both the REG and proposed Bayesian methods noticeably reduce this bias and produce estimates much closer to the true finite-population mean. As ρ increases, the value of σ decreases, and the outcome becomes more strongly associated with the covariates. Correspondingly, the values of PSD, CV, and NSE decrease as ρ increases. Comparing the two scenarios, the smaller sample size scenario $(n_1, n_2) = (400, 100)$ yields larger PSD, CV, and NSE values than $(n_1, n_2) = (800, 200)$, reflecting the increased uncertainty associated with smaller samples.

Table 4. Posterior summaries for the finite-population mean under different values of ρ and two sample-size scenarios.

ρ	Method	$(n_1, n_2) = (800, 200)$				$(n_1, n_2) = (400, 100)$			
		PM	PSD	CV	NSE	PM	PSD	CV	NSE
0.2	Naive	30.1	0.225	0.007	0.007	30.1	0.319	0.011	0.010
	REG	28.9	0.256	0.009	0.008	28.9	0.374	0.013	0.012
	Proposed	28.9	0.440	0.015	0.015	28.9	0.649	0.022	0.024
0.3	Naive	30.2	0.147	0.005	0.005	30.3	0.207	0.007	0.007
	REG	28.9	0.163	0.006	0.005	28.9	0.237	0.008	0.007
	Proposed	28.8	0.278	0.010	0.010	28.8	0.410	0.014	0.015
0.5	Naive	30.1	0.085	0.003	0.003	30.2	0.120	0.004	0.004
	REG	28.9	0.089	0.003	0.003	28.9	0.130	0.005	0.004
	Proposed	28.9	0.153	0.005	0.005	28.9	0.225	0.008	0.008
0.7	Naive	30.2	0.059	0.002	0.002	30.2	0.083	0.003	0.003
	REG	28.9	0.053	0.002	0.002	28.9	0.077	0.003	0.002
	Proposed	28.9	0.092	0.003	0.003	28.9	0.135	0.005	0.005
0.8	Naive	30.1	0.050	0.002	0.002	30.2	0.070	0.002	0.002
	REG	28.9	0.039	0.001	0.001	28.9	0.057	0.002	0.002
	Proposed	28.9	0.067	0.002	0.002	29.0	0.099	0.003	0.004

Table 5 reports the simulation performance measures under the two sample size scenarios. The naive estimator has the largest ARB and PRMSE across all settings, and its coverage probability is essentially zero. The REG and proposed Bayesian methods considerably improve both point estimation accuracy and interval estimation performance. The proposed Bayesian method generally achieves smaller ARB and PRMSE values than the REG estimator and consistently produces larger coverage probabilities, although this improvement comes at the cost of wider interval estimates. Similar patterns are observed under both sample size scenarios. The smaller sample size scenario $(n_1, n_2) = (400, 100)$ generally produces larger ARB, PRMSE, and interval widths, reflecting the increased uncertainty associated with smaller samples.

The REG estimator is computationally simpler and generally produces narrower interval estimates. In contrast, the proposed Bayesian approach jointly models latent NPS weights and missing PS outcomes, allowing uncertainty in these quantities to be incorporated into posterior inference. These results illustrate the practical trade-off between interval width and coverage.

For the 95% HPD intervals, the quantity $\sigma / \sqrt{N}$ plays a crucial role. Because N is much larger than σ^2 , the resulting HPD intervals are expected to be relatively narrow. This makes it more difficult for the true finite-population mean to remain within the HPD intervals across simulation runs. Furthermore, as ρ increases, the value of σ decreases, leading to even narrower intervals. Therefore, coverage probabilities may be smaller than the nominal 95% level. Very large values of ρ , such as $\rho = 0.8$, may also be less realistic in practice because they imply very small residual variability.

In addition, the Bayesian credible intervals are constructed under the Bayesian paradigm, whereas the empirical coverage probabilities are evaluated from a frequentist repeated-sampling perspective. Under the frequentist paradigm, the parameter is fixed and the data are random, whereas under the Bayesian paradigm, the parameter is treated as random conditional on the observed data. Consequently, lower frequentist coverage probabilities do not necessarily imply poor Bayesian inference. Furthermore, the use of informative priors for w_1 and y_2 may contribute to relatively concentrated posterior intervals. Less informative priors with larger variances may improve the empirical coverage probabilities.

Table 5. Simulation performance measures under different values of ρ and two sample-size scenarios.

ρ	Method	$(n_1, n_2) = (800, 200)$				$(n_1, n_2) = (400, 100)$			
		ARB	PRMSE	Cov	Wid	ARB	PRMSE	Cov	Wid
0.2	Naive	0.043	1.270	0.000	0.873	0.045	1.330	0.011	1.230
	REG	0.028	0.881	0.362	0.992	0.040	1.270	0.395	1.450
	Proposed	0.019	0.749	0.781	1.700	0.028	1.100	0.784	2.510
0.3	Naive	0.046	1.330	0.000	0.570	0.048	1.390	0.000	0.803
	REG	0.017	0.548	0.387	0.630	0.025	0.795	0.376	0.918
	Proposed	0.012	0.480	0.773	1.080	0.019	0.722	0.754	1.590
0.5	Naive	0.042	1.210	0.000	0.328	0.043	1.250	0.000	0.464
	REG	0.010	0.320	0.345	0.346	0.015	0.463	0.338	0.504
	Proposed	0.007	0.282	0.707	0.591	0.011	0.415	0.715	0.873
0.7	Naive	0.043	1.240	0.000	0.229	0.044	1.280	0.000	0.323
	REG	0.007	0.229	0.293	0.206	0.010	0.311	0.332	0.299
	Proposed	0.006	0.206	0.590	0.356	0.008	0.291	0.600	0.525
0.8	Naive	0.042	1.210	0.000	0.192	0.043	1.250	0.000	0.271
	REG	0.006	0.189	0.289	0.152	0.009	0.273	0.245	0.221
	Proposed	0.005	0.179	0.498	0.260	0.008	0.258	0.492	0.383

The BMI application required approximately 19 s on a standard desktop computer. The simulation studies were conducted on the WPI Turing HPC cluster. For the two sample-size scenarios considered in this paper, the total runtimes were approximately 9.3 h for $(n_1, n_2) = (400, 100)$ and 12.7 h for $(n_1, n_2) = (800, 200)$. As expected, computational cost increases with sample size because the proposed method jointly updates latent NPS weights and missing PS outcomes within the Gibbs sampler. Nevertheless, the method remains computationally feasible for the scenarios considered in this study.

7. Discussion

This paper proposes a Bayesian framework for integrating non-probability samples and probability samples for finite-population inference. The method combines outcome information from the non-probability sample with survey-weight information from the probability sample through a hierarchical Bayesian model. The adjusted non-probability sample weights and the missing probability sample outcomes are treated as unknown quantities within the posterior computation.

A simulation study was conducted to compare the proposed Bayesian method with the naive estimator and the REG estimator under different levels of association between the covariates and the outcome variable. The naive estimator produced substantial bias because it ignores the informative sampling mechanism in the non-probability sample. Both the REG and proposed Bayesian methods reduced this bias. In general, the proposed Bayesian method produced wider posterior intervals because it accounts for uncertainty in the latent weights and missing outcomes.

The proposed method was also applied to a BMI dataset. The naive estimator overestimated the finite-population mean, whereas the REG and proposed Bayesian methods produced estimates closer to the weighted population summaries computed from the full dataset. The weighted population median was also reported because the population mean may be sensitive to extreme BMI values.

The simulation study and BMI application suggest that jointly modeling latent weights and missing outcomes can reduce bias arising from informative non-probability sampling. In the BMI application, the outcome distribution exhibited mild right-skewness and several extreme observations. This indicates that the normality assumption adopted in the current model may be restrictive in some practical settings. More flexible Bayesian models that accommodate skewness or heavy tails may therefore improve finite-population inference. Another important challenge in non-probability surveys is measurement error in the study variable. Sen and Lahiri (2026) [19] showed that improving both representativeness and measurement accuracy can lead to better statistical inference. Extending the proposed Bayesian framework to account for measurement error is another interesting direction for future research.

Future work may also consider Bayesian semiparametric extensions based on mixture models and stick-breaking process priors. Related Bayesian nonparametric approaches include the Dirichlet process of Ferguson (1973) [20], Bayesian mixture models of Escobar and West (1995) [21], and stick-breaking constructions of Ishwaran and James (2001) [22]. Extensions to more complex survey designs and hierarchical population structures also warrant further investigation.

Author Contributions: Conceptualization, Q.C. and B.N.; methodology, Q.C. and B.N.; software, Q.C.; validation, Q.C. and B.N.; formal analysis, Q.C. and B.N.; investigation, Q.C. and B.N.; data curation, Q.C.; writing—original draft preparation, Q.C.; writing—review and editing, Q.C. and B.N.; visualization, Q.C.; supervision, B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The simulation data were generated within the study. The BMI data used in this study are not publicly available due to privacy considerations. Additional information and code are available from the corresponding author upon reasonable request.

Acknowledgments: Results in this paper were obtained in part using a high-performance computing system acquired through NSF MRI grant DMS-1337943 to WPI.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Notation Summary

Table A1. Notation summary.

Symbol	Description
N	Population size
n_1	NPS sample size
n_2	PS sample size
$\hat{n}_1$	Effective NPS sample size used for calibration
$\hat{n}_2$	Effective PS sample size used for calibration
W_{1i}	Original survey weight for unit i in the NPS, unobserved
W_{2i}	Original survey weight for unit i in the PS, observed
w_{1i}	Adjusted survey weight for unit i in the NPS, unobserved
w_{2i}	Adjusted survey weight for unit i in the PS, observed
y_{1i}	Outcome for unit i in the NPS, observed
y_{2i}	Outcome for unit i in the PS, unobserved
$\bar{Y}$	Finite population mean, target parameter
β	Regression coefficient vector in the outcome model, unknown parameter
$\tilde{\sigma}^2$	Outcome variance parameter, unknown parameter

Appendix B. Derivation of σ in the Simulation Design

Recall that

$$Y_i = \eta_i + e_i, \quad e_i \sim (1-p)N(0, \gamma\sigma^2) + pN(0, \sigma^2)$$

and e_i is independent of η_i .

Using the definition of covariance,

$$\text{Cov}(\eta_i, Y_i) = E(\eta_i Y_i) - E(\eta_i)E(Y_i).$$

Since $Y_i = \eta_i + e_i$,

$$E(\eta_i Y_i) = E\{\eta_i(\eta_i + e_i)\} = E(\eta_i^2) + E(\eta_i e_i).$$

Because e_i is independent of η_i and $E(e_i) = 0$,

$$E(\eta_i e_i) = E(\eta_i)E(e_i) = 0 \quad \text{and} \quad E(Y_i) = E(\eta_i + e_i) = E(\eta_i).$$

Therefore,

$$\text{Cov}(\eta_i, Y_i) = E(\eta_i^2) - [E(\eta_i)]^2 = \text{Var}(\eta_i).$$

Similarly,

$$\text{Var}(Y_i) = \text{Var}(\eta_i + e_i) = \text{Var}(\eta_i) + \text{Var}(e_i) = \text{Var}(\eta_i) + (1-p)\gamma\sigma^2 + p\sigma^2.$$

Hence,

$$\rho = \text{Corr}(\eta_i, Y_i) = \frac{\text{Cov}(\eta_i, Y_i)}{\sqrt{\text{Var}(\eta_i) \text{Var}(Y_i)}} = \frac{\text{Var}(\eta_i)}{\sqrt{\text{Var}(\eta_i)(\text{Var}(\eta_i) + (1-p)\gamma\sigma^2 + p\sigma^2)}}.$$

Squaring both sides gives

$$\rho^2 = \frac{\text{Var}(\eta_i)}{\text{Var}(\eta_i) + (1-p)\gamma\sigma^2 + p\sigma^2}.$$

Solving for σ ,

$$\sigma = \sqrt{\frac{\text{Var}(\eta_i)}{(1-p)\gamma + p} \frac{1 - \rho^2}{\rho^2}}.$$

References

1. Elliott, M.R.; Valliant, R. Inference for nonprobability samples. *Stat. Sci.* **2017**, *32*, 249–264. [\[CrossRef\]](#)
2. Beaumont, J.-F. Are probability surveys bound to disappear for the production of official statistics? *Surv. Methodol.* **2020**, *46*, 1–28.
3. Beaumont, J.-F.; Rao, J.N.K. Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *Surv. Stat.* **2021**, *83*, 11–22.
4. Chen, Y.; Li, P.; Wu, C. Doubly robust inference with nonprobability survey samples. *J. Am. Stat. Assoc.* **2020**, *115*, 2011–2021.
5. Rao, J.N.K. On making valid inferences by integrating data from surveys and other sources. *Sankhyā Ser. B* **2021**, *83*, 242–272. [\[CrossRef\]](#)
6. Kim, J.; Park, S.; Chen, Y.; Wu, C. Combining non-probability and probability survey samples through mass imputation. *J. R. Stat. Soc. Ser. A* **2021**, *184*, 941–963. [\[CrossRef\]](#)
7. Kim, J.K.; Morikawa, K. An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Stat. Assoc. Bull.* **2023**, *75*, 8–27. [\[CrossRef\]](#)
8. Marella, D. Adjusting for selection bias in non-probability samples by empirical likelihood approach. *J. Off. Stat.* **2023**, *39*, 151–172. [\[CrossRef\]](#)
9. Sakshaug, J.W.; Wisniowski, A.; Ruiz, D.A.P.; Blom, A.G. Supplementing small probability samples with nonprobability samples: A Bayesian approach. *J. Off. Stat.* **2019**, *35*, 653–681. [\[CrossRef\]](#)
10. Salvatore, C.; Biffignandi, S.; Sakshaug, J.W.; Wisniowski, A.; Struminskaya, B. Bayesian integration of probability and non-probability samples for logistic regression. *J. Surv. Stat. Methodol.* **2023**, *12*, 458–492. [\[CrossRef\]](#)
11. Nandram, B.; Rao, J.N.K. A Bayesian approach for integrating a small probability sample with a non-probability sample. In Proceedings of the American Statistical Association, Virtual, 8–12 August 2021; pp. 1568–1603.
12. Nandram, B.; Rao, J.N.K. Bayesian predictive inference when integrating a nonprobability sample and a probability sample. *arXiv* **2023**, arXiv:2305.08997.
13. Wu, C. Statistical Inference with non-probability survey samples. *Surv. Methodol.* **2022**, *48*, 283–311.
14. Nandram, B. Bayesian predictive inference under informative sampling via surrogate samples. In *Bayesian Statistics and Its Applications*; Upadhyay, S.K., Singh, U., Dey, D.K., Eds.; Anamaya: New Delhi, India, 2007; Chapter 25, pp. 356–374.
15. Nandram, B.; Rao, J.N.K. Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Stat. Appl.* **2024**, *22*, 343–374.
16. Nandram, B. Bayesian small area estimation using a non-probability sample. In Proceedings of the Special Proceedings, Society of Statistics, Computing and Application, Rajasthan, India, 26–28 February 2024; pp. 1–30.
17. Lockwood, A.; Nandram, B. Bayesian predictive inference of a finite population mean without specifying the relation between the study variable and the covariates. *Surv. Methodol.* **2024**, *50*, 465–486.
18. Kish, L. *Survey Sampling*; John Wiley & Sons: New York, NY, USA, 1965.
19. Sen, A.; Lahiri, P. Improving measurement error and representativeness in nonprobability surveys. *Surv. Methodol.* **2026**, *52*, 1–26.
20. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1973**, *1*, 209–230. [\[CrossRef\]](#)
21. Escobar, M.D.; West, M. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* **1995**, *90*, 577–588. [\[CrossRef\]](#)
22. Ishwaran, H.; James, L.F. Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* **2001**, *96*, 161–173. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.