Electronic Journal of Statistics
Vol. 13 (2019) 1608–1645
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1538

Fully Bayesian estimation under informative sampling

Luis G. León-Novelo[*] **and Terrance D. Savitsky**[†]

1200 Pressler St. Suite E805, Houston, TX, 77030, USA and
PSB Suite 1950, 2 Massachusetts Avenue, NE Washington, DC 20212-0001, USA.
e-mail: Luis.G.LeonNovelo@uth.tmc.edu; Savitsky.Terrance@bls.gov

Abstract: Survey data are often collected under informative sampling designs where subject inclusion probabilities are designed to be correlated with the response variable of interest. The data modeler seeks to estimate the parameters of a population model they specify from these data. Sampling weights constructed from marginal inclusion probabilities are typically used to form an exponentiated pseudo likelihood as a plug-in estimator in a partially Bayesian pseudo posterior. We introduce the first fully Bayesian alternative, based on a Bayes rule construction, that simultaneously performs weight smoothing and estimates the population model parameters in a construction that treats the response variable(s) and inclusion probabilities as jointly randomly generated from a population distribution. We formulate conditions on known marginal and pairwise inclusion probabilities that define a class of sampling designs where L_1 consistency of the joint posterior is guaranteed. We compare performances between the two approaches on synthetic data. We demonstrate that the credibility intervals under our fully Bayesian method achieve nominal coverage. We apply our method to data from the National Health and Nutrition Examination Survey to explore the relationship between caffeine consumption and systolic blood pressure.

Keywords and phrases: Bayesian penalized B-splines, informative sampling, inclusion probabilities, NHANES, sampling weights, survey sampling.

Received July 2018.

Contents

---

[*] Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston- School of Public Health

[†] Office of Survey Methods Research, Bureau of Labor Statistics

1. Introduction

Our focus is on inference about a population generating distribution, P_0, estimated on an observed sample data acquired under a complex survey sampling design from a finite population. A known sampling design distribution, P_ν, defines a joint distribution over the random inclusions of units from the finite population (previously generated from P_0). The sampling design distribution, P_ν, often intentionally constructs a correlation between specified unit inclusion probabilities and the response variable of interest; for example, the U.S. Bureau of Labor Statistics (BLS) employs proportion-to-size (PPS) sampling design in the Current Employment Statistics (CES) survey. The CES is administered with the purpose to construct total employment statistics for area and industry-indexed domains. Establishments with larger employment values are assigned higher unit inclusion probabilities since these larger establishments drive more of the variance in total employment point estimates for domains. Survey designs that instantiate a correlation between unit inclusions and the response variable of interest are termed, "informative", because the inclusion probabilities are informed by the response variable. The balance of information in the observed sample is different from that in the population under informative sampling, such that naively estimating a proposed population model, P, on the observed sample will result in biased inference about P_0, that is supposed to govern the finite population (Savitsky and Toth, 2016). Let N denote the size of a finite population with variables drawn under P_0 and n denote a sample taken from that population, governed by P_ν. Sampling weights constructed to be inversely proportional to unit inclusion probabilities, π_i $i = 1, \ldots, n$, may be used, along with the response variable(s) of interest, y_i, to perform inference about P_0 from the observed sample.

Our contribution is to develop the *first* fully Bayesian estimator that supposes both (y_i, π_i) are generated from the population-generating distribution,

P_0, while P_ν governs the taking of all possible samples from the finite population. Let $\delta_\ell \in \{0, 1\}$ be a random variable that indexes the inclusion of unit ℓ, $\ell \in (1, \ldots, N)$ from the population into the observed sample. P_ν governs δ_ℓ, where $\pi_\ell = Pr(\delta_\ell = 1 \mid y_\ell)$, for $\ell \in (1, \ldots, N)$. The π_ℓ in the CES are constructed from the y_ℓ, where the survey design assigns higher inclusion probabilities with larger employment. The implication is that both (y_ℓ, π_ℓ) are both generated (randomly) under P_0. So we construct a joint *population* likelihood for $(y_\ell, \pi_\ell)_{\ell=1,\ldots,N}$ and it is the definition of a joint distribution that makes our approach fully Bayesian as contrasted with a plug-in approach that treats the π_ℓ as fixed. We introduce a Bayes rule procedure in the sequel that constructs a joint likelihood for the *observed sample*, $(y_i, \pi_i)_{i=1,\ldots,n}$, from the specified population joint likelihood. We treat the π_ℓ as random by assigning a conditional likelihood, $p\left(\pi_\ell \mid y_\ell, \boldsymbol{\kappa}\right)$. Our approach of treating π_ℓ as random (with respect to P_0) allows us to incorporate dependencies induced by the sampling design, P_ν. We are the first to demonstrate that utilizing a fully Bayesian approach leads to credibility intervals that contract on valid frequentist confidence intervals, unlike the usual plug-in approaches. We further demonstrate the frequentist consistency of our fully Bayesian estimator in the sequel that accounts for joint distribution, (P_0, P_ν), over population generation and the taking of an informative sample. In summary, this paper lays down the foundational concepts and implementation procedure for how to construct and utilize a fully Bayesian estimator as an alternative to the plug-in estimator developed by (Savitsky and Toth, 2016).

1.1. Review of methods to account for informative sampling

There are three broad classes of approaches to adjust model estimation on the observed sample acquired under an informative sampling (IS) design for inference about P_0. One approach parameterizes the sampling design into the model estimated on the sampled data (Little, 2004). The sampling design, itself, is often a nuisance to the data analyst, however, and their focus on the parameters of P_0 requires them to marginalize over parameters indexing the sampling design distribution. It is also often the case that the analyst does not know the sampling design to parameterize it.

The next two classes of modeling adjustments employ a sampling weight that is constructed to be inversely proportional to the marginal inclusion probability, $\pi_i = P\left(\delta_i = 1\right)$, for each unit, $i \in \{1, \ldots, n\}$, where n denotes the number of units in the observed sample, S. The inclusion of unit, i, from the population, U, in the sample, $S \subset U$ is indexed by the random variables, $\delta_i \in \{0, 1\}$, governed by P_ν. The likelihood contribution for each unit in the observed sample is adjusted by its associated sampling weight, such that the joint adjusted likelihood over the sample provides an approximation to the balance of information in the finite population.

The second class of modeling approaches employs a particular form for the likelihood (Dong et al., 2014; Kunihama et al., 2016; Rao and Wu, 2010; Si et al., 2015), rather than allowing the analyst to specify a population model, P, because the inferential focus is not on parameters of the generating model, P_0, but on domain-level estimation of simple mean and total statistics, such as total employment for a geographic area and industry combination.

In the third class of modeling approaches, Savitsky and Toth (2016) construct a sampling-weighted pseudo posterior distribution by exponentiating each unit likelihood contribution, under the analyst-specified model, by its sampling weight, to produce, $p\left(y_i|\delta_i = 1, \boldsymbol{\lambda}\right)^{w_i}$. Exponentiating by the sampling weight, $w_i \propto 1/\pi_i$, constructs the pseudo likelihood used to estimate the pseudo posterior when convolved with the prior distributions for model parameters, $\boldsymbol{\lambda}$. Savitsky and Toth (2016) demonstrate that estimation of (the parameters of) P_0 from the pseudo posterior distribution is asymptotically unbiased. This approach provides a "plug-in" approximation to the population likelihood (for n observations), in that the sampling inclusion probabilities, (π_i), are assumed fixed. On the one hand, one may suppose that a sequential generation for the finite population under P_0 and the subsequent taking of the sample under P_ν (where the two distributions are convolved under informative sampling). In this case, one may view the information contained in the generated finite population as fixed. Inclusion probabilities are assigned by P_ν to the units comprising the population, and so, are also viewed as fixed (given the underlying finite population). From a Bayesian perspective, on the other hand, there is a dynamic process of population generation and the taking of an informative sample. If the population changes, then sample inclusion probabilities, which depend on the population response values, will also change.

The absence of an established procedure to incorporate the inclusion probabilities into the likelihood of a particular population model chosen by the data analyst hinders the use of Bayesian statistics in the analysis of survey data released with sampling weights. To address this lack of a fully Bayesian treatment for modeling data under informative sampling, this paper formulates a fully Bayesian set-up to jointly model (y_i, π_i) from P on the observed sample. Our approach *extends* the set-up of Pfeffermann et al. (1998) to a fully Bayesian formulation by specifying a conditional population model, $p(\pi_i \mid y_i, \boldsymbol{\kappa})$, for the inclusion probabilities, $(\pi_i)_{i \in U}$. We then apply a Bayes rule approach to construct, $p(y_i, \pi_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}, \delta_i = 1)$, that conditions on the observed sample, $(\delta_\ell = 1)_{\ell=1,\ldots,N}$, where N denotes the population size, $|U|$, and $\mathbf{x}_i$ denotes a vector of covariates. We employ this full likelihood defined on the observed sample to estimate the joint posterior distribution for model parameters, $(\boldsymbol{\theta}, \boldsymbol{\kappa})$. We demonstrate in the sequel that the informativeness of the sampling design improves estimation efficiency for our fully Bayes approach that jointly models (y_i, π_i) as compared to the plug-in, pseudo posterior framework (that treats π_i as fixed). We formulate conditions that guarantee the L_1 frequentist consistency of our fully Bayes estimator. We further conduct a simulation study that demonstrates the credibility sets generated under our fully Bayes construction achieve nominal coverage (and we also demonstrate, by contrast, that credibility sets of the sampling-weighted, pseudo posterior distribution do not achieve nominal coverage).

Pfeffermann et al. (1998) focus on maximum likelihood point estimation from $p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}, \delta_i = 1)$, rather than the joint distribution over the response and inclusion probabilities, as do we. We also relax their theoretical condition of independence among the sampled units that they employ to guarantee consistency of the point estimate to asymptotic independence in our result for the L_1 consistency of our fully Bayes posterior distribution.

Pfeffermann et al. (2006) also extend Pfeffermann et al. (1998) to a partially Bayesian estimation, but they treat $(\pi_\ell)_{\ell=1,\ldots,N}$ as *fixed*; in particular, they formulate a likelihood for the observed sample using Equation 3.3 of Pfeffermann et al. (1998) that uses Bayes rule to achieve $p(y_i|x_i, \boldsymbol{\theta}, \boldsymbol{\kappa}, \delta_i = 1) = [\mathbb{E}(\pi_i \mid y_i, x_i, \boldsymbol{\kappa})/\mathbb{E}(\pi_i \mid x_i, \boldsymbol{\theta}, \boldsymbol{\kappa})] \times p(y_i \mid \theta)$, where they only specify the population distribution, $p(y_i \mid \theta)$ and do not treat π_i as random. Instead, they specify a point estimate, $\mathbb{E}(\pi_i|\ldots)$ (obtained without specifying distribution for π_i). By contrast, our formulation for the joint likelihood takes the view that the inclusion probabilities, (π_i), are generated along with the population, $(y_i, \mathbf{x}_i)_{i=1,\ldots,N}$, $N = |U|$. So the approach of Pfeffermann et al. (2006) may be viewed to be *not* fully Bayesian as is ours because it does not specify a joint model for $p(\pi_i|y_i, \boldsymbol{\kappa})$.

We derive the joint likelihood for the observed sample, $p(y_i, \pi_i|\delta_i = 1, \cdots)$, that adapts a Bayes rule approach from Pfeffermann et al. (1998), which adjusts the analyst-specified population likelihood, $p(y_i, \pi_i|\cdots)$, from which we formulate our joint posterior distribution for the observed sample in Section 2. We provide examples that demonstrate how to implement the fully Bayes estimator and elucidate its properties. Conditions are constructed that guarantee a frequentist L_1 contraction of our fully Bayesian posterior distribution, P, onto the true generating distribution in Section 3. This section can be skipped without loss of understanding the practical application of the proposed method. We conduct a simulation study under a synthetically generated population from a known, P_0, from which we draw samples under an informative, proportion-to-size (PPS) sampling design, to compare the performances of our fully Bayesian estimator and the pseudo posterior in Section 4. We reveal that: (1) fully Bayes credible intervals achieve nominal frequentist coverage while the pseudo posterior under covers; and (2) the fully Bayes point estimates are robust, in terms of bias and mean square error (MSE), against high variability of inclusion probabilities in contrast to the pseudo posterior. We illustrate our fully Bayesian estimator by assessing the relationship between caffeine consumption and systolic blood pressure using NHANES data in Section 5. The paper concludes with a discussion in Section 6. The proofs for the main result, along with two enabling results and other materials are contained in the Appendix.

2. Fully Bayesian estimator under Bayes rule

We proceed to formulate a posterior distribution for the *observed* sample taken under an informative design as a function of inclusion probabilities to produce unbiased inference and correct uncertainty quantification for the population model. Our approach first constructs a *joint* distribution for the response, y, and inclusion probabilities, π, for the *population*, which we define as a fully Bayesian specification (because we treat π as random). We then use Bayes rule to adjust the joint population likelihood to an expression for the observed sample by incorporating conditional distributions for inclusion indicators, ($\delta_i \in \{0,1\}$). The adjusted likelihood is unbiased for the population model parameters estimated on the sample. We later show the novel result that credibility sets of the posterior distribution estimated from the fully Bayes adjusted likelihood achieve nominal coverage.

From Bayes rule,

$$p(y_i, \pi_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}, \delta_i = 1) = \frac{\Pr(\delta_i = 1 | y_i, \pi_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) \times p(y_i, \pi_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa})}{\Pr(\delta_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa})}, \tag{1}$$

where the expression on the left-hand side of Equation (1) adjusts the joint *population* likelihood, $p(y_i, \pi_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa})$, specified on the right-hand side for the realized sample. This joint population likelihood is defined with respect to P_{λ_0}, where $\lambda = (\boldsymbol{\theta}, \boldsymbol{\kappa})$ and λ_0 is the true generating value. So we treat both (y, π) as jointly generated from P_{λ_0}. The δ_i are governed by the sampling design distribution, P_ν, that we will formally introduce in Section 3.

We may simplify the numerator by plugging in (See also equation (7.1) in Pfeffermann et al., 1998),

$$\Pr(\delta_i = 1 | y_i, \pi_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) = \pi_i. \tag{2}$$

We next compute the double expectation with respect to $(\boldsymbol{\kappa}, \boldsymbol{\theta})$ in the denominator,

$$\begin{aligned}
\Pr(\delta_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) &= \int \int \Pr(\delta_i = 1 | y_i, \pi_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) p(y_i, \pi_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) \, d\pi_i dy_i \\
&= \int \int \pi_i p(\pi_i | y_i, \mathbf{x}_i, \boldsymbol{\kappa}) \, d\pi_i \, p(y_i | \mathbf{x}_i, \boldsymbol{\theta}) \, dy_i \\
&= \int E(\pi_i | y_i, \mathbf{x}_i, \boldsymbol{\kappa}) \, p(y_i | \mathbf{x}_i, \boldsymbol{\theta}) \, dy_i \\
&= E_{y_i | \mathbf{x}_i, \boldsymbol{\theta}} \left[E(\pi_i | y_i, \mathbf{x}_i, \boldsymbol{\kappa}) \right],
\end{aligned} \tag{3}$$

where we assume separability of the parameter sets, $\boldsymbol{\kappa}, \boldsymbol{\theta}$, in the conditional distributions for $\pi_i | y_i, \mathbf{x}_i$ and $y_i | \mathbf{x}_i$,

$$\begin{aligned}
p(y_i, \pi_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) &= p(\pi_i | y_i, \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) p(y_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) \\
&= p(\pi_i | y_i, \mathbf{x}_i, \boldsymbol{\kappa}) p(y_i | \mathbf{x}_i, \boldsymbol{\theta}),
\end{aligned} \tag{4}$$

where $p(\pi_i | y_i, \mathbf{x}_i, \boldsymbol{\kappa})$ and $p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$ denote conditional distributions for the population. We make note that we have factorized the joint population distribution over (y_i, π_i), such that information about the sampling design is parameterized in the conditional model for $\pi_i | y_i$, rather than in the conditional distribution for $y_i | x_i$.

We define $p_s(y_i, \pi_i|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) = p(y_i, \pi_i|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}, \delta_i = 1)$ (where subscript, s, indexes the observed sample of size, n) to be the likelihood contribution for each unit, $i \in S$, and plug in Equations (2) and (3) into Equation (1) to obtain,

$$p_s(y_i, \pi_i|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) = \frac{\pi_i p(\pi_i|y_i, \mathbf{x}_i, \boldsymbol{\kappa})}{E_{y_i'|\mathbf{x}_i,\boldsymbol{\theta}}\left[E(\pi_i'|y_i', \mathbf{x}_i, \boldsymbol{\kappa})\right]} \times p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \tag{5}$$

Let $N = |U|$ denote the population size and $S = \{S(1), \ldots, S(n)\} := \{i \in \{1, \ldots, N\} : \delta_i = 1\}$, the population indices of individuals or units included in the random sample S of size $n = |S|$. Therefore, $\Pr(S = s) \propto \prod_{i=1}^{n} \pi_{s(i)}$. Let the generic vector $\mathbf{r}^{(s)} := \{r_1^{(s)}, \ldots, r_n^{(s)}\} := \{r_{s(1)}, \ldots, r_{s(n)}\}$ represent the values of variable, r, observed in the sample. With this notation the likelihood over the observed sample of size, n, is specified by,

$$\ell(\boldsymbol{\theta}, \boldsymbol{\kappa}; \mathbf{y}^{(s)}, \boldsymbol{\pi}^{(s)}, \mathbf{x}^{(s)}) = \prod_{i=1}^{n} \left[p_s(y_i^{(s)}, \pi_i^{(s)} \mid \mathbf{x_i}^{(s)}, \boldsymbol{\theta}, \boldsymbol{\kappa})\right]$$

from which we may incorporate the prior distributions to formulate the posterior distribution observed on the sample,

$$p_s(\boldsymbol{\theta}, \boldsymbol{\kappa} \mid \mathbf{y}^{(s)}, \boldsymbol{\pi}^{(s)}, \mathbf{x}^{(s)}) \propto \ \ell(\boldsymbol{\theta}, \boldsymbol{\kappa}; \mathbf{y}^{(s)}, \boldsymbol{\pi}^{(s)}, \mathbf{x}^{(s)}) \times \text{Prior}(\boldsymbol{\theta}) \times \text{Prior}(\boldsymbol{\kappa}).$$

We have now constructed the backbone of a fully Bayesian model that takes into account informative sampling designs for estimation of population model parameters. Our result jointly models the response and the inclusion probabilities, *i.e.*, (y_i, π_i), using only quantities observed in the sample; in particular, the joint distribution of (y_i, π_i) are different in the observed sample and in the population, and we have corrected for this difference in a way that allows us to make unbiased estimation of the parameters of the population model. In contrast to Si et al. (2015), we do not impute the values for the non sampled units. By assuming a population distribution for the inclusion probabilities, the variability of the inclusion probabilities is modeled and the noise not depending on the response (and $\mathbf{x}_i$) is discarded; more specifically $p_s(y_i, \pi_i \mid \cdots)$ in Equation (5) is constructed from a conditional model that regresses π_i on y_i and uses the expected value in the denominator.

Achieving our result of Equation (5) requires specification of a joint population generating distribution, $p(y_i, \pi_i|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\kappa})$. Specifying a population model for the response, $p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$, is commonly done. Going a step further, as we do, to incorporate the inclusion probabilities in the population model, is not commonly done. By contrast, the pseudo posterior approach, which is formally introduced in the sequel, considers the inclusions probabilities (or equivalently, the sampling weights) as fixed. Under an informative sampling design where the (π_i) are constructed to depend on the (y_i), each time we generate new values of the response variable we are prompted to update the inclusion probabilities for all units in the population after fixing a sampling design. So we believe the treatment of π_i as random is natural as they are conditioned on the values of population variables.

The price the modeler pays for this fully Bayesian approach is that they are required to specify a conditional distribution of the inclusion probabilities for all units in the population, $p(\pi_i \mid y_i, \boldsymbol{\kappa})$. This requirement raises two computational difficulties: Firstly, even if the specified population likelihood, $p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$, and the associated prior, $p(\boldsymbol{\theta})$, yield a closed form conditional posterior distribution, this construction for the population model does not produce a conjugate posterior under our formulation to correct for informative sampling (IS) because the sampling design is not ignorable; Secondly, the computation of the expected value in the denominator of Equation (5) is a computational bottleneck as we are required to compute this expected value for each observation in every Gibbs sampler iteration. We cope with the first difficulty by relying on the black box solver Stan (Carpenter et al., 2016), that uses a "no U turn" version of Hamiltonian Monte Carlo (HMC) approach to draw samples from the full joint posterior distribution, which is a version of a Metropolis-Hastings sampler that produces proposals which partially suppress the random walk. HMC performs well under (is insensitive to) non-conjugacy, but may be sensitive to varying scale in the data. Other alternative sampling approaches include the generalized elliptical slice sampler of Nishihara et al. (2014), which is another version of an MH sampler that automatically generates proposals that is not sensitive to the scale of the data, and may be embedded in a Gibbs scan to draw samples from a block of correlated parameters. Regarding the second computational challenge, we proceed to construct a set of conditions on $p(\pi_i \mid y_i, \mathbf{x}_i, \boldsymbol{\kappa})$ and the likelihood that guarantee the availability of a closed form expression for this expected value.

We next specify a class of conditional population distributions from Equation (4) that yield a closed form result for the expectation step in denominator of (5), which simplifies posterior computation. Let $\mathbf{v}_i$ and $\mathbf{u}_i$ be subvectors of $\mathbf{x}_i$, the covariates used to specify the conditional distribution of $\pi_i \mid y_i, \mathbf{x}_i, \boldsymbol{\kappa}$ and $y_i \mid \mathbf{x}_i, \boldsymbol{\theta}$, respectively; that is, $\pi_i \mid y_i, \mathbf{x}_i, \boldsymbol{\kappa} \sim \pi_i \mid y_i, \mathbf{v}_i, \boldsymbol{\kappa}$ and $y_i \mid \mathbf{x}_i, \boldsymbol{\theta} \sim y_i \mid \mathbf{u}_i, \boldsymbol{\theta}$. Note that we allow for $\mathbf{v}_i$ and $\mathbf{u}_i$ to have common covariates. Let $\text{normal}(x \mid \mu, s^2)$ denote the normal distribution pdf with mean μ and variance s^2 evaluated at x, and $\text{lognormal}(\cdot \mid \mu, s^2)$ denote the lognormal pdf, so that $X \sim \text{lognormal}(\mu, s^2)$ is equivalent to $\log X \sim \text{normal}(\mu, s^2)$. Going forward, we assume that π_i is proportional, as opposed to exactly equal, to the inclusion probability for unit i. In other words, no restriction is imposed on $\sum_i \pi_i$ where the index i could run over the population or sample indices. We will see in the sequel that normalizing the sum of the inverse inclusion probabilities *is* required to regulate the estimation of posterior uncertainty under the pseudo likelihood approach, but such is not required for our fully Bayesian formulation.

Theorem 1. *If* $p(\pi_i \mid y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = \text{lognormal}(\pi_i \mid h(y_i, \mathbf{v}_i, \boldsymbol{\kappa}), \sigma_{\pi^2})$, *with the function* $h(y_i, \mathbf{v}_i, \boldsymbol{\kappa})$ *of the form* $h(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) + t(\mathbf{v}_i, \boldsymbol{\kappa})$ *where* $\sigma_\pi^2 = \sigma_\pi^2(\boldsymbol{\kappa}, \mathbf{v}_i)$, *possibly a function of* $(\boldsymbol{\kappa}, \mathbf{v}_i)$ *then*

$$p_s(y_i, \pi_i \mid \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) = \frac{\text{normal}\left(\log \pi_i \mid g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) + t(\mathbf{v}_i, \boldsymbol{\kappa}), \sigma_\pi^2\right)}{\exp\left\{t(\mathbf{v}_i, \boldsymbol{\kappa}) + \sigma_\pi^2/2\right\} \times M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta})} \times p(y_i \mid \mathbf{u}_i, \boldsymbol{\theta})$$

with $M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}) := E_{y|\mathbf{u}_i, \boldsymbol{\theta}}\left[\exp\left\{g(y_i, \mathbf{v}_i, \boldsymbol{\kappa})\right\}\right]$.

The proof of this theorem Appendix Section A. If both M_y and $p(y_i \mid \cdots)$ admit closed form expressions, then $p_s(y_i, \pi_i \mid \cdots)$ has a closed form, as well; for example, when $g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = \kappa_y y_i$ with $\kappa_y \in \boldsymbol{\kappa} \in \mathbb{R}$, then $M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta})$ is the moment generating function (MGF) of $y_i \mid \boldsymbol{\theta}$ evaluated at κ_y, which will have a closed form defined on $\mathbb{R}$ for typically-used normal, binomial and probit population models for the responses. The closed form for $M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta})$ implies a closed form for $p_s(y_i, \pi_i \mid \cdots)$. Analogously, we may consider an interaction between y and $\mathbf{v}$, using $g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = (\kappa_y + \mathbf{v}_i^t \boldsymbol{\kappa}_{\mathbf{v}}) y_i \equiv t y_i$ with $\boldsymbol{\kappa} = (\kappa_y, \boldsymbol{\kappa}_{\mathbf{v}})$. In this case, we achieve, $M_y(t; \cdots)$, which is the MGF evaluated at t.

By selecting a lognormal distribution for $\pi_i \mid y_i$ we achieve an expectation with respect to y of the exponential of a linear κy term in y, which is just the moment generating function for y, parameterized by κ. This important result outlines a class of population models where the integration (expectation) step in the denominator of Equation (5) admits a closed-form expression for easy posterior sampling. In a fully Bayesian implementation, the integration step must be evaluated on every posterior sampling iteration. So our development of Theorem 1 to highlight the class of models with closed forms is an important development. We conduct further discussion in Section 6 of on-going work to explore the use of numerical integration to allow a wider and more general class of models for a fully Bayesian implementation.

Our assumption of a lognormal distribution for $\pi_i | y_i$ is also mathematically appealing since the inclusion probability, π_i, for individual, i, is usually composed from the product of inclusion probabilities of selection across the stages of the multistage survey design. If each of these stagewise probabilities are lognormal then their product, π_i, is lognormal as well. We next apply Theorem 1 to some common settings.

2.1. Linear regression population model

Assume the linear regression model for the population, $p(y_i \mid \mathbf{u}_i, \boldsymbol{\theta})$, is constructed as,

$$y_i \mid \mathbf{u}_i, \boldsymbol{\theta} \sim \text{normal}\left(\mathbf{u}_i^t \boldsymbol{\beta}, \sigma_y^2\right), \quad \text{with } \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_y^2) \tag{6}$$

and the conditional population model for inclusion probabilities is specified as,

$$\pi_i \mid y_i, \mathbf{v}_i, \boldsymbol{\kappa} \sim \text{lognormal}\Big(\kappa_y y_i + \mathbf{v}_i^t \boldsymbol{\kappa}_x, \sigma_\pi^2\Big), \quad \text{with } \boldsymbol{\kappa} = (\kappa_y, \boldsymbol{\kappa}_x, \sigma_\pi^2) \tag{7}$$

This construction results from setting, $g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = k_y y_i$, $t(\mathbf{v}_i, \boldsymbol{\kappa}) = \mathbf{v}_i^t \boldsymbol{\kappa}_x$, $\sigma_\pi^2(\boldsymbol{\kappa}, \mathbf{v}_i) = \sigma_\pi^2$. Here $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}_x$ are vectors of regression coefficients that include an intercept, so the first entry of both $\mathbf{u}_i$ and $\mathbf{v}_i$ equals 1. We select prior distributions,

$$\begin{aligned} \boldsymbol{\beta} \sim& \text{MVN}(\mathbf{0}, 100\mathbf{I}) \\ (\kappa_y, \boldsymbol{\kappa}_x) \sim& \text{MVN}(\mathbf{0}, 100\mathbf{I}) \end{aligned} \tag{8}$$

$$\sigma_y, \sigma_\pi \overset{\text{iid}}{\sim} \text{Cauchy}^+(0,1)$$

where $\text{Cauchy}^+(m, v)$ denotes the Cauchy distribution with location and scale parameters m and v, respectively, restricted to the positive real line; $\text{MVN}(\mathbf{m}, \Sigma)$ denotes the multivariate normal distribution with mean vector $\mathbf{m}$ and variance-covariance matrix Σ; and $\mathbf{I}$ the identity matrix.

Since $y \sim \text{normal}(m, s^2)$ admits a closed form expression for $M_y(t) = \exp(tm + t^2s^2/2)$, we apply Theorem 1 to obtain,

$$p_s(y_i, \pi_i \mid \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) = \frac{\text{normal}(\log \pi \mid \kappa_y y_i + \mathbf{v}_i^t \boldsymbol{\kappa}_x, \sigma_\pi^2)}{\exp\left\{\mathbf{v}_i^t \boldsymbol{\kappa}_x + \sigma_\pi^2/2 + \kappa_y \mathbf{u}_i^t \boldsymbol{\beta} + \kappa_y^2 \sigma_y^2/2\right\}} \times \text{normal}(y_i \mid \mathbf{u}_i^t \boldsymbol{\beta}, \sigma_y^2) \tag{9}$$

The resulting form of the expression for $\log p_s(\dots)$ is provided in an Appendix, Section B.

2.2. Other population models

Since a closed-form moment generating function of $y_i \mid \mathbf{u}_i, \boldsymbol{\theta}$ defined in $\mathbb{R}$ implies a closed form for $p_s(y_i, \pi_i \mid \dots)$, extensions for the logistic, probit and Poisson regressions to perform asymptotically unbiased estimations on the observed sample are straightforward; for example, employing the conditional lognormal model, Equation (7), for the inclusion probabilities, but now selecting a probit construction for a dichotomous, $y_i \sim \text{Bernoulli}(p_i)$, with $p_i = \Phi^{-1}(\mathbf{x}_i^t \boldsymbol{\theta})$, where Φ^{-1} is the inverse of the standard normal CDF. We note the closed form MGF, $E(e^{ty}) = (1 - p + pe^t)$ for all t, and obtain,

$$p_s(y_i, \pi_i \mid \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) = \frac{\text{normal}(\log \pi_i \mid \kappa_y y_i + \mathbf{v}_i^t \boldsymbol{\kappa}_x, \sigma_\pi^2)}{\exp\left\{\mathbf{v}_i^t \boldsymbol{\kappa}_x + \sigma_\pi^2/2\right\}(1 - p_i + p_i e^{\kappa_y})} \times \text{Bernoulli}(y \mid p_i)$$

with $\text{Bernoulli}(y \mid p_i) = p_i^{y_i}(1 - p_i)^{1 - y_i}$.

2.3. Splines basis population model

We extend the linear regression model to a splines setting for situations where the relationship between the response and the predictor is not linear. Here, $p(y_i \mid \mathbf{u}_i, \boldsymbol{\theta})$ is constructed as,

$$y_i \sim \text{normal}(\mu(\mathbf{u}_i), \sigma_y^2) \quad \text{with } \mu(\mathbf{u}_i) := B(\mathbf{u}_i)\boldsymbol{\beta} \tag{10}$$

with $B(\mathbf{u}_i) := (B_{i1}, \dots, B_{ib})$, the vector of penalized B-spline coefficients associated with $\mathbf{u}_i$, where b denotes the number of spline bases. We construct a rank-deficient multivariate Gaussian prior for the coefficients, $\boldsymbol{\beta}$, which penalizes or regulates the smoothness of the estimated function, μ, by employing a $b \times b$ matrix, $Q = D^t D$, where D is the discretized kth difference operator (Speckman and Sun, 2003) in the density kernel,

$$\sigma_{\boldsymbol{\beta}}^{-(b-k)/2} \exp -\frac{1}{2\sigma_{\boldsymbol{\beta}}^2} \boldsymbol{\beta}^t Q \boldsymbol{\beta}.$$

The order, k, regulates the resulting smoothness of the fitted function or its order of differentiability. $B(\mathbf{u}_i)$ has at most k entries different from zero. The penalized pdf specified for $\boldsymbol{\beta}$ is not proper since the dimension of $\boldsymbol{\beta}$ is b, but Q is of rank $b-k$. We observe that the MLE for $(\boldsymbol{\beta}, \sigma_y^2, \sigma_{\boldsymbol{\beta}}^2)$ is a type of ridge regression estimator since it satisfies,

$$(B^t B/\hat{\sigma}_y^2 + Q/\hat{\sigma}_{\boldsymbol{\beta}}^2)\hat{\boldsymbol{\beta}} = B^t \mathbf{y}/\hat{\sigma}_y^2.$$

Our simulation scenarios in Section 4 will utilize this penalized B-spline formulation with $(b = 8, k = 4)$. The priors for model variances are set as weakly informative proper priors,

$$\sigma_y \sim \text{normal}^+(0, 10^2), \quad \sigma_{\boldsymbol{\beta}} \sim \text{Cauchy}^+(0, 10) \quad \text{and } \sigma_\pi \sim \text{Cauchy}^+(0, 1);$$

where $\text{normal}^+(m, s^2)$ denotes the normal distribution with mean and variance parameters m and s^2, respectively, restricted to be positive. $\tau_{\boldsymbol{\beta}} = 1/\sigma_{\boldsymbol{\beta}}^2$ is referred to as a complexity parameter.

Since the spline model is a particular case of the linear regression model in Subsection 2.1, $p_s(y_i, \pi_i \mid \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}, \boldsymbol{\kappa})$, with $\boldsymbol{\theta} := (\boldsymbol{\beta}, \sigma_y^2)$, for the spline model has the same expression as that for the linear model in Equation (9) after replacing $\mathbf{u}_i^T \boldsymbol{\beta}$ for $B_i \boldsymbol{\beta}$ (assuming our continued employment of a lognormal prior for π_i in Equation (7)).

We may further generalize the populations for y_i and $\pi_i | y_i$ with the addition of latent random effects (for example, to parameterize the dependence induced by the clustering step of a multistage sampling design). We defer such models to on-going research in order to focus on comparing general properties of the fully Bayes method to the pseudo posterior. See Section 6 for more discussion of future research.

2.4. Estimation of population distribution of inclusion probabilities

For a simple one-stage design the whole population of inclusion probabilities or equivalent is available, because their values are required to draw a sample. For a multistage cluster sample like a survey of individuals within households, we draw a sample recursively, only specifying inclusion probabilities for units selected at the previous level. This procedure provides sampling probabilities only for those lowest-level units or individuals that are selected into the sample. It may be useful for the survey administrator to estimate the population distribution of

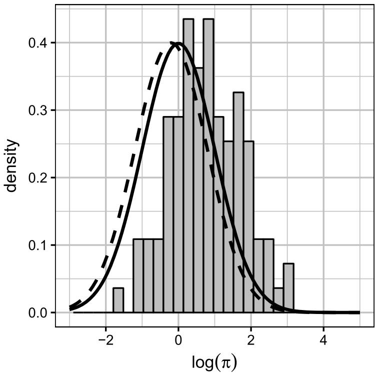

FIG 1. *Histogram of the $n = 100$ log-transformed sampled inclusion probabilities along with the true simulated standardized normal population density (solid line) and estimated normalized population density (dashed line) with mean and variance estimated using Equation (11).*

inclusion probabilities from the observed sample in order to verify a subgroup of units targeted for over-sampling do, indeed, express relatively higher inclusion probabilities.

We select a model that satisfies the conditions of Theorem 1, but set $y_i \equiv 0$. This is, $g(y_i, \mathbf{v}_i, \boldsymbol{\kappa})$ is constant and without loss of generality equal to zero. Then, $M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}) = 1$ and Theorem 1 implies

$$p_s(\pi_i \mid \mathbf{v}_i, \boldsymbol{\kappa}) = \frac{\text{normal}\left(\log \pi_i \mid t(\mathbf{v}_i, \boldsymbol{\kappa}), \sigma_\pi^2\right)}{\exp\left\{t(\mathbf{v}_i, \boldsymbol{\kappa}) + \sigma_\pi^2/2\right\}} \tag{11}$$

If we assume $\pi_i \sim \text{lognormal}(t(\mathbf{v}_i, \boldsymbol{\kappa}), \sigma_\pi^2)$ for $i = 1, \ldots, N$ with $\Pr(\delta_i = 1) = \pi_i / \sum_{i=1}^N \pi_i$, we can make inference about the parameters $\boldsymbol{\kappa}$ and σ_π^2 using only the observed inclusion probabilities. As an example we simulate $\pi_1, \ldots, \pi_{10^5} \overset{\text{iid}}{\sim} \text{lognormal}(\boldsymbol{\kappa} = 0, \sigma_\pi^2 = 1)$. (Our choice of the lognormal distribution implies $t(\mathbf{v}_i, \boldsymbol{\kappa}) \equiv \boldsymbol{\kappa}$.) We draw $n = 100$ inclusion probabilities (π_i) under IS, and estimate the parameter values $\boldsymbol{\kappa}$ and σ_π^2 via Equation (11) using the priors outlined in Equation (8). Figure 1 shows the histogram of the log transformed sampled inclusion probabilities observed in our sample of $n = 100$, along with the standardized density of generated inclusion probabilities for our simulated population, displayed in the solid line, and our estimated normal density with population mean $E(\boldsymbol{\kappa} \mid \pi_1, \ldots, \pi_n)$ and variance $E(\sigma_\pi^2 \mid \pi_1, \ldots, \pi_n)$, displayed in the dashed line. Notice that we employ the unstandardized sampling weights. If the sampling weights are multiplied by a constant $c > 0$; e.g., $c := n / \sum_{i=1}^n \pi_i$, we are then estimating $\boldsymbol{\kappa} + \log c$.

2.5. Pseudolikelihood alternative to the fully Bayes estimator

The pseudo posterior approach is a plug-in method that is not fully Bayesian, which we will use to compare with the fully Bayes method we propose in this paper. The pseudo likelihood is formed by exponentiating the likelihood contribution for each observed unit in the sample by its sampling weight, $w_i^{(s)}$, in,

$$\text{pseudo}(\boldsymbol{\theta}; \mathbf{y}^{(s)}, \mathbf{w}^{(s)}) = \prod_{i=1}^{n} p(y_i^{(s)} \mid \mathbf{x}^{(s)}{}_i, \boldsymbol{\theta})^{w_i^{(s)}}$$

where $w_i^{(s)} \propto 1/\pi_i^{(s)}$ is the sampling weight that is standardized such that $\sum w_i^{(s)} = n$, where n denotes to the observed sample size, which is required to regulate the amount of estimated posterior uncertainty. Let $\mathbf{w}^{(s)} := (w_1^{(s)}, \ldots, w_n^{(s)})$ denote the vector of these unit-indexed sampling weights that are typically published with the survey data in order to correct for IS. Inference is carried out by constructing the pseudo posterior distribution be proportional to $\text{pseudo}(\boldsymbol{\theta}; \mathbf{y}^{(s)}, \mathbf{w}^{(s)}) \times \text{Prior}(\boldsymbol{\theta})$ (Savitsky and Toth, 2016). Although the pseudo likelihood is an improper distribution, it admits a proper joint posterior distribution under employment of proper priors. Classical implementations of the sampling weighted pseudo likelihood, by contrast to the Bayesian implementation, employ MLE estimation of population model parameters on the observed IS. The standard error of the estimate of parameter, θ, is estimated via resampling methods such as balanced repeated replication (McCarthy, 1969; Krewski and Rao, 1981) or Fray's method (Judkins, 1990) when the sampling design and the information provided to the analyst allow it.

The advantages of the plug-in pseudo posterior method over the proposed fully Bayesian construction are: (i) Estimation does not require a custom posterior sampler. Relatively slight modifications are performed to the population model sampler to incorporate the pseudo likelihood; (ii) Specification of $\pi_i \mid y_i, \cdots$ for the population is not required; (iii) There is no expected value in Equation (3) to compute as there is in the fully Bayes method. The disadvantages of the pseudo posterior approach are: (i) It is not fully Bayesian; (ii) It does not discard variation in weights that do not depend on the response, such that noise (unrelated to the response) is introduced into the estimation of the pseudo posterior distribution; (iii) The weights must be normalized to regulate the amount of estimated posterior uncertainty, which is not required for the fully Bayes approach. We will show in the sequel that the fully Bayes approach expresses superior performance in estimation of posterior uncertainty than does the pseudo posterior.

3. Posterior consistency

We proceed to construct conditions on the population generating model and sample inclusion probabilities that guarantee the contraction of our fully Bayes estimator to the true joint generating distribution, P_{λ_0}. The conditions constructed for the sample inclusion probabilities define a class of sampling designs for which we would expect asymptotically unbiased estimation of our population model parameters on the observed data. Our approach will be to create a novel estimator / likelihood for our adjusted likelihood in Equation (5) that allows to apply the approach of Savitsky and Toth (2016) in formulating and proving the frequentist consistency result.

3.1. Set-up

Let $\nu \in \mathbb{Z}^+$ index a sequence of finite populations, $\{U_\nu\}_{\nu=1,\ldots,N_\nu}$, each of size, $|U_\nu| = N_\nu$, such that $N_\nu < N_{\nu'}$, for $\nu < \nu^{'}$, so that the finite population size grows as ν increases. We employ subscript, ν, because the process of rendering a sample involves the generation of a population followed by the assignment of sample inclusion probabilities and the taking of a sample from that population using the sampling weights. The entire sequence is repeated on each increment of ν. A sampling design is defined by placing a distribution on a vector of inclusion indicators, $\boldsymbol{\delta}_\nu = (\delta_{\nu 1}, \ldots, \delta_{\nu N_\nu})$, linked to the units comprising the population, U_ν, that we use to draw a sample of size $n_\nu < N_\nu$. We construct a sampling distribution by specifying marginal inclusion probabilities, $\pi_{\nu i} = \Pr\{\delta_{\nu i} = 1 | Y_{\nu,i} = y_{\nu i}\}$ for all $i \in U_\nu$ and the second-order pairwise probabilities, $\pi_{\nu ij} = \Pr\{\delta_{\nu i} = 1 \cap \delta_{\nu j} = 1 | Y_{\nu,i} = y_{\nu i}, Y_{\nu,j} = y_{\nu j}\}$ for $i,j \in U_\nu$, where $Y_{\nu,1}, \ldots, Y_{\nu,N_\nu}$ index random variables that are independently distributed according to some unknown distribution P_θ (with density, p_θ) defined on the sample space, $(\mathcal{Y}_\nu, \mathcal{A}_\nu)$.

In the usual survey sampling set-up, the inclusion probabilities are treated as *fixed*. Under the Bayesian paradigm, however, Y_ν is random variable drawn from a super-population, such that if a new population, U_ν, with associated observed values $(y_{\nu i})$, is drawn, the marginal inclusion probabilities, $(\pi_{\nu i})$, will update. So we condition the inclusion probabilities explicitly on the population response values for emphasis, unlike in Savitsky and Toth (2016). We envision both inclusion probabilities, $(\pi_{\nu i})$, and our data, $(y_{\nu,i})$, as generated from a super-population under some true model. We construct the joint density of the model for $(y_{\nu i}, \pi_{\nu i})$ for each $i = 1, \ldots, N_\nu$ as,

$$p_\lambda\left(y_{\nu i}, \pi_{\nu i}\right) = p_\theta\left(y_{\nu i}\right) \times p_\kappa\left(\pi_{\nu i} | Y_{\nu i} = y_{\nu i}\right), \tag{12}$$

where $\lambda = (\theta, \kappa)$ denotes our model parameters.

We use the Bayes rule construction we earlier illustrated to formulate a sampling design-adjusted fully Bayesian joint likelihood for the unknown population value, λ_0, which we suppose generates the finite population values that we observe in our sample, $(y_{\nu,i}, \pi_{\nu,i})_{i \in U_\nu : \delta_{\nu i} = 1}$,

$$\begin{aligned}
& p^\pi_\lambda\left(x_{\nu i} \delta_{\nu i}\right) && (13)\\
&\quad := \left[\frac{\pi_{\nu i} \times p_\lambda\left(y_{\nu i}, \pi_{\nu i}\right)}{\mathbb{E}_\theta\left\{\Pr_\kappa\left(\delta_{\nu i} = 1 | Y_{\nu i} = y_{\nu i}\right)\right\}}\right]^{\delta_{\nu i}} && (14)\\
&\quad := \left[\frac{\pi_{\nu_i}}{\mathbb{E}_\theta\left(\pi^\kappa_{\nu i}\right)} \times p_\lambda\left(y_{\nu i}, \pi_{\nu i}\right)\right]^{\delta_{\nu i}} && (15)\\
&\quad := \left[\frac{\pi_{\nu_i}}{\pi^\lambda_{\nu i}}\, p_\lambda\left(y_{\nu i}, \pi_{\nu i}\right)\right]^{\delta_{\nu i}}, \quad i \in U_\nu, && (16)
\end{aligned}$$

where we collect, $x_{\nu i} = (y_{\nu i}, \pi_{\nu i})$,
and define $\pi^{\kappa}_{\nu i} := \text{Pr}_{\kappa}\left(\delta_{\nu i} = 1 | Y_{\nu i} = y_{\nu i}\right) = \mathbb{E}_{\kappa}\left(\pi_{\nu i} | Y_{\nu i} = y_{\nu i}\right)$, where the second equality is derived in Pfeffermann et al. (1998). We further define $\pi^{\lambda}_{\nu i} := \mathbb{E}_{\theta}\left(\pi^{\kappa}_{\nu i}\right) = \mathbb{E}_{\theta}\left\{\mathbb{E}_{\kappa}\left(\pi_{\nu i} | Y_{\nu i} = y_{\nu i}\right)\right\}$ for ease-of-reading. Our construction of our joint likelihood in Equation (16) will allow us to follow the approach of Savitsky and Toth (2016) in our main result after adding a new condition for the consistency of $\pi^{\lambda}_{\nu i}$ in Section 3.2.

We conduct Bayesian inference under the likelihood of Equation (16) by assigning prior, Π, on the parameter space, Λ, such that $\lambda_0 \in \Lambda$, which produces the sampling design-adjusted posterior mass,

$$\Pi^{\pi}\left(B | x_{\nu 1}\delta_{\nu 1}, \ldots, x_{\nu N_{\nu}}\delta_{\nu N_{\nu}}\right) = \frac{\int_{\lambda \in B} \prod_{i=1}^{N_{\nu}} \frac{p^{\pi}_{\lambda}}{p^{\pi}_{\lambda_0}}(x_{\nu i}\delta_{\nu i}) d\Pi(\lambda)}{\int_{\lambda \in \Lambda} \prod_{i=1}^{N_{\nu}} \frac{p^{\pi}_{\lambda}}{p^{\pi}_{\lambda_0}}(x_{\nu i}\delta_{\nu i}) d\Pi(\lambda)}, \tag{17}$$

that we use to formulate our theoretical result. We define $p^{\pi}_{\lambda_0}(x_{\nu i}\delta_{\nu i}) = p^{\pi}_{\lambda_0}(x_{\nu i})^{\delta_{\nu i}}$, that confines evaluation to the observed sample.

We will utilize the empirical distribution construction to establish a bound used to prove our result,

$$\mathbb{P}^{\pi}_{N_{\nu}} = \frac{1}{N_v} \sum_{i=1}^{N_{\nu}} \frac{\delta_{\nu i}}{\pi^{\lambda}_{\nu i}} \delta\left(X_i\right), \tag{18}$$

which is nearly identical to Savitsky and Toth (2016), only now the denominator is a modeled quantity dependent on λ. We construct the associated expectation functional, $\mathbb{P}^{\pi}_{N_{\nu}} f = \frac{1}{N_v} \sum_{i=1}^{N_{\nu}} \frac{\delta_{\nu i}}{\pi^{\lambda}_{\nu i}} f\left(X_i\right)$. We make a similar adjustment to the Hellinger distance,
$d^{\pi,2}_{N_{\nu}}\left(p_{\lambda_1}, p_{\lambda_2}\right) = \frac{1}{N_{\nu}} \sum_{i=1}^{N_{\nu}} \frac{\delta_{\nu i}}{\pi^{\lambda}_{\nu i}} d^2\left(p_{\lambda_1}(x_{\nu i}), p_{\lambda_2}(x_{\nu i})\right)$,
where $d\left(p_{\lambda_1}, p_{\lambda_2}\right) = \left[\int \left(\sqrt{p_{\lambda_1}} - \sqrt{p_{\lambda_2}}\right)^2 d\mu\right]^{\frac{1}{2}}$ (for dominating measure, μ).

3.2. *Main result*

We outline the 6 conditions and our result for completeness, though they are very similar to Savitsky and Toth (2016) save for the replacement of $\pi_{\nu i}$ with the modeled $\pi^{\lambda}_{\nu i}$. The first 3 conditions place restrictions on the generating distribution and prior, while the following three do the same for the sampling design distribution. Suppose we have a sequence, $\xi_{N_{\nu}} \downarrow 0$ and $N_{\nu}\xi^2_{N_{\nu}} \uparrow \infty$ and $n_{\nu}\xi^2_{N_{\nu}} \uparrow \infty$ as $\nu \in \mathbb{Z}^+ \uparrow \infty$ and any constant, $C > 0$.

(A1) (Local entropy condition - Size of model)

$$\sup_{\xi > \xi_{N_{\nu}}} \log N\left(\xi/36, \{\lambda \in \Lambda_{N_{\nu}} : d_{N_{\nu}}\left(P_{\lambda}, P_{\lambda_0}\right) < \xi\}, d_{N_{\nu}}\right) \leq N_{\nu}\xi^2_{N_{\nu}},$$

(A2) (Size of space)

$$\Pi\left(\Lambda \backslash \Lambda_{N_{\nu}}\right) \leq \exp\left(-N_{\nu}\xi^2_{N_{\nu}}\left(2(1 + 2C)\right)\right)$$

(A3) (Prior mass covering the truth)

$$\Pi\left(P_{\lambda}:\left(-\mathbb{E}_{\lambda_0}\log\frac{p_{\lambda}}{p_{\lambda_0}}\leq\xi_{N_{\nu}}^2\right)\cap\left(\mathbb{E}_{\lambda_0}\left[\log\frac{p_{\lambda}}{p_{\lambda_0}}\right]^2\leq\xi_{N_{\nu}}^2\right)\right)\geq\exp\left(-N_{\nu}\xi_{N_{\nu}}^2C\right)$$

(A4) (Non-zero Inclusion Probabilities)

$$\sup_{\nu}\left[\frac{1}{\min\limits_{i\in U_{\nu}}\pi_{\nu i}^{\lambda}}\right]\leq\gamma,\text{ with }P_{\lambda_0}\text{−probability 1.}$$

(A5) (Asymptotic Independence Condition)

$$\limsup_{\nu\uparrow\infty}\max_{i\neq j\in U_{\nu}}\left|\frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}}-1\right|=\mathcal{O}(N_{\nu}^{-1}),\text{ with }P_{\lambda_0}\text{−probability 1}$$

such that $\pi_{\nu ij}$ factors to $\pi_{\nu i}\pi_{\nu j}$ for N_{ν} sufficiently large where there exists some $C_3>0$,

$$N_{\nu}\sup_{\nu}\max_{i\neq j\in U_{\nu}}\left[\frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}}-1\right]\leq C_3.$$

(A6) (Constant Sampling fraction) For some constant, $f\in(0,1)$, that we term the “sampling fraction”,

$$\limsup_{\nu}\left|\frac{n_{\nu}}{N_{\nu}}-f\right|=\mathcal{O}(1),\text{ with }P_0\text{−probability 1.}$$

(A7) (Convergence of the Point Estimate)

$$\limsup_{\nu\uparrow\infty}\max_{i\in U_{\nu}}\left|\mathbb{E}_{\lambda}\left[\pi_{\nu i}\right]-\mathbb{E}_{\lambda_0}\left[\pi_{\nu i}\right]\right|=\mathcal{O}(N_{\nu}^{-1}),\text{ with }P_0\text{−probability 1.}$$

Condition (A1) restricts the growth in the size of the model space (Ghosal et al., 2000) by bounding the growth in the logarithm of the covering number, $N\left(\xi/36,\{\lambda\in\Lambda_{N_{\nu}}:d_{N_{\nu}}\left(P_{\lambda},P_{\lambda_0}\right)\right)$, defined as the *minimum* number of balls of radius $\xi/36$ needed to cover $\{P\in\mathcal{P}_{N_{\nu}}:d_{N_{\nu}}\left(P,P_0\right)<\xi\}$ under distance metric, $d_{N_{\nu}}$. Condition (A2) allows, but restricts, the prior mass placed on the uncountable portion of the model space, such that we may direct our inference to an approximating sieve, $\mathcal{P}_{N_{\nu}}$. This sequence of spaces “trims” away a portion of the space that is not entropy bounded (in condition (A1)). Condition (A3) ensures the prior, Π, assigns mass to convex balls in the vicinity of P_{λ_0}.

The next three conditions, together, restrict the class of sampling designs under which our result is guaranteed. Condition (A4) requires the sampling design to assign a positive probability for inclusion of every unit in the population because the restriction bounds the sampling inclusion probabilities away from 0. Since the maximum inclusion probability is 1, the bound, $\gamma\geq 1$. Unlike in Savitsky and Toth (2016), however, $\pi_{\nu i}^{\lambda}$ is a smoothed model estimator under our fully Bayes construction, which discards the variation in $\pi_{\nu i}$ that is unrelated to $y_{\nu i}$, such that the value of γ, in practice, is expected to be lower than in Savitsky and Toth (2016). Condition (A5) restricts the result to sampling designs where the dependence among lowest-level sampled units attenuates to 0 as $\nu \uparrow \infty$. Dependence in a multistage design is driven the higher level sampling stages; for example, PSUs. Since the number of PSUs increases in the limit of N_ν, this condition is not very restrictive and admits nearly all sampling designs used, in practice. Condition (A6) ensures that the observed sample size, n_ν, limits to ∞ along with the size of the partially-observed finite population, N_ν. The denominator of our Bayes rule posterior estimator of Equation (16) is a conditional expectation with respect to our model, P_λ. We require convergence of this point estimate in order to achieve the bound specified in condition (A5). Condition (A7) is not needed in Savitsky and Toth (2016), since the inclusion probabilities, (π_ν), are assumed fixed.

Theorem 2. *Suppose conditions (A1)-(A7) hold. Then for sets* $\Lambda_{N_\nu} \subset \Lambda$, *constants,* $K > 0$, *and* M *sufficiently large,*

$$\mathbb{E}_{P_{\lambda_0},P_\nu} \Pi^\pi \left(\lambda : d^\pi_{N_\nu}\left(P_\lambda, P_{\lambda_0}\right) \geq M\xi_{N_\nu} | x_{\nu 1}\delta_{\nu 1}, \ldots, x_{\nu N_\nu}\delta_{\nu N_\nu}\right) \leq \frac{16\gamma^2\left[\gamma + C_3\right]}{\left(Kf + 1 - 2\gamma\right)^2 N_\nu \xi^2_{N_\nu}} + 5\gamma^2 \exp\left(-\frac{K n_\nu \xi^2_{N_\nu}}{2\gamma}\right), \tag{19}$$

which tends to 0 *as* $(n_\nu, N_\nu) \uparrow \infty$.

The rate of convergence is injured for a sampling distribution, P_ν, that assigns relatively low inclusion probabilities to some units in the finite population such that γ will be relatively larger. Our result differs from Savitsky and Toth (2016) in that 5γ that multiplies the second term is replaced by $5\gamma^2$, here, though the value γ is expected to be substantially lower, as earlier discussed, for our modeled estimate, $\pi^\lambda_{\nu i}$, than for the raw, $\pi_{\nu i}$. In practice, γ will be much lower for the fully Bayes method than for the pseudo posterior due to the weight smoothing (in the modeled $\pi^\lambda_{\nu i}$), such that convergence will be faster. We illustrate this point in the simulation study for the spline model in Section 4. Similarly, the larger the dependence among the finite population unit inclusions induced by P_ν, the higher will be C_3 and the slower will be the rate of contraction.

The proof of our main result proceeds by bounding the numerator of Equation (17) (on the set $\left\{\lambda : d^\pi_{N_\nu}\left(P_\lambda, P_{\lambda_0}\right) \geq M\xi_{N_\nu}\right\}$), from above, and the denominator, from below, and is the same as that outlined in Savitsky and Toth (2016). So we do not repeat it here. The proof relies two enabling results, which express some differences in this work, however, from Savitsky and Toth (2016). So those Lemma are reformulated and proved in an Appendix section C.

4. Simulation study

We explore the bias and coverage performance of our fully Bayes formulation and compare it to the plug-in pseudo posterior model under employment of an informative, probability proportional-to-size (PPS) sampling design; first in a simple linear regression and later in a nonlinear, spline model setting. We construct a synthetic population in each iteration of our Monte Carlo simulation, from which we draw two samples: 1. An informative sample (IS) taken under the PPS design; 2. A simple random sample (SRS). The IS is analyzed using both the fully Bayes and the pseudo posterior methods. The SRS is estimated using our population model with no needed correction. The SRS serves as a something of a gold standard comparator to assess the efficacy of bias removal and uncertainty estimations using the methods that correct for informative sampling, which are estimated on the IS. Of course, an informative sampling design can be more efficient (lower variance) than an SRS, though an IS is often constructed as much for convenience (and cost) as for efficiency. We estimate bias, mean square error (MSE), coverage of central 95% credible interval (CI) and its average length by repeating this this process $M = 1,000$ times in a Monte Carlo framework. The three analysis methods (that we label, "Full" (fully Bayesian), "Pseudo" (Pseudo Posterior) and "SRS" (simple random sampling)) will use the same population distribution for $y_i \mid \boldsymbol{\theta}, \ldots$ and prior for $\boldsymbol{\theta}$. The true value of the generic simulation model parameter, η, is denoted by η^{TRUE}.

We conclude our simulation study with a test of the robustness of the lognormal population model specification for $\pi|y$ in a real world setting by treating our observed NHANES sample as a population and drawing samples from it using the inverse of published unit sampling weights as inclusion probabilities. Since the multistage NHANES sampling design is constructed algorithmically, we don't expect the resulting distribution for $\pi|y$ to be lognormal.

4.1. Simple linear regression (SLR): PPS design

We focus on estimating the population generating for the intercept coefficient, β_0, specified for the linear regression framework in Section 2.1. The fully Bayes estimation model assumes a lognormal conditional likelihood for $\pi_i \mid y_i, \cdots$, which is right skewed. We explore three simulation scenarios; In the first two, the true generating distribution of the inclusion probabilities is a gamma distribution and, therefore, skewed as in the the fully Bayes analysis model. In the third simulation scenario, we explore the robustness of the fully Bayes model to this assumption by generating the inclusion probabilities from a (symmetric) beta distribution. The difference between the first and second scenarios is that we vary the rate hyperparameter of the gamma distribution to induce more variance (and skewness) into the generated (π_i) in the second scenario. To explore the effect of not correcting for IS, we also analyze the informative sample with the population model ignoring the weights (label"Pop").

4.1.1. Scenarios SLR: π-skewed.

We call the first two simulation scenarios "SLR: π-skewed with low variance" and "SLR: π-skewed with high variance". Both set $\pi_i \overset{\text{iid}}{\sim} \text{gamma}(2, rate = b_\pi^{\text{TRUE}})$ (See 1 (b) below) where $b_\pi^{\text{TRUE}} = 2$, in the former, and $b_\pi^{\text{TRUE}} = 1$, in the latter.

We set $M = 10^3$ to be the total number of synthetic population datasets or Monte Carlo iterations, and $N = 10^5$ to be the number of individuals / units in the population, and we set $n = 10^3$ to be the sample size.
Let $\boldsymbol{\theta}^{\text{TRUE}} = (\beta_0^{\text{TRUE}}, \beta_1^{\text{TRUE}}, \beta_2^{\text{TRUE}}, \sigma_y^{2,\text{TRUE}}) = (0, 1, 1, 0.1^2)$ under the following simulation procedure:

- For $m = 1, \ldots, M$, Monte Carlo iterations
  1. Generate the population, *i.e.*, for $i = 1, \ldots, N$,
     (a) Draw $\mathbf{u}_i \overset{\text{iid}}{\sim} \text{uniform}(0, 1)$
     (b) Draw $\pi_i \overset{\text{iid}}{\sim} \text{gamma}(2, \mathit{rate} = b_\pi^{\text{TRUE}})$ (which produces unnormalized π_i).
     (c) Generate the population response:
     $$y_i \mid \mathbf{u}_i, \pi_i, \boldsymbol{\theta}^{\text{TRUE}} \sim \text{normal}(\beta_1^{\text{TRUE}} \mathbf{u}_i + \beta_2^{\text{TRUE}} \pi_i, \sigma_y^{2,\text{TRUE}})$$
  2. Draw two samples of size $n = 10^3$:
     (a) Take an IS with, $\Pr[(y_i, \pi_i) \in \text{sample}] = \pi_i / \sum_{i'=1}^{N} \pi_{i'}$.
     (b) Take an SRS sample.
  3. Conduct estimation using the priors given in Equation (8).
     (a) The informative sample is estimated using,
        i. Full: the fully Bayes linear regression approach outlined in Subsection 2.1 with parameters, $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_y^2)$, $\boldsymbol{\kappa} = (\kappa_y, \sigma_\pi^2)$, where we note there is no $\mathbf{v}$ or $\boldsymbol{\kappa}_x$.
        ii. Pseudo: the pseudo posterior linear regression formulation.
        iii. Pop: linear regression ignoring informative sampling.
     (b) SRS: The SRS is estimated using the linear regression population model of Equation (6).
  4. (a) Store the posterior expected values of β_0 and β_1 under the analyses 3(a) i-iii and 3(b).
     (b) Also compute the central 95% credible interval (CI) for β_0 and β_1 under the three analyses and store their lengths and indicator of whether they contain the true intercept coefficient under the SLR model, $\beta_0^{SLR} := \beta_0^{\text{TRUE}} + E(\pi_i)$ (see below) and slope β_1^{TRUE}.
- Using the values stored in step 4, (a) estimate the Bias and MSE of the estimate of the true intercept coefficient under the SLR model and, (b), estimate the coverage and average length of the central 95% credible interval (CI) for β_0^{SLR}. Do the same for the estimate of β_1^{TRUE}.

Note that the synthetic data generating likelihood, $\pi_i \mid y_i, \mathbf{v}_i, \boldsymbol{\kappa}$, is *not* the lognormal distribution for the population estimation model; that is, the fully Bayes (estimation) model is misspecified. Also note that the fully Bayes population model in (6) is misspecified under the three analyses. More specifically, the expected value $y_i \mid \mathbf{u}_i, \boldsymbol{\theta}$, after integrating out π_i, matches that corresponding

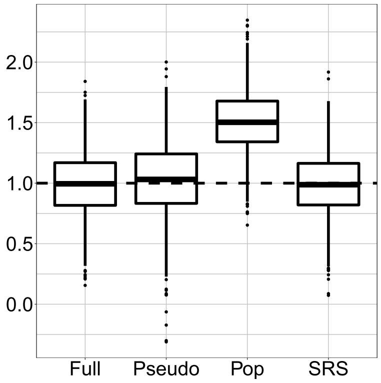

Fig 2. *Estimation of intercept coefficient under SLR: π skewed with low variance. Here, $\pi_i \overset{\text{iid}}{\sim}$ gamma$(2, b_\pi^{TRUE} = 2)$ in step 1(a) in Section 4.1.1. Box plots of posterior expected values of β_0 are displayed under IS for the fully Bayes, pseudo posterior approaches, ignoring the informative design (Pop), and under SRS. The horizontal dashed line represents simulation value of the intercept under the SLR model, i.e., $\beta_0^{SLR} = \beta_0^{TRUE} + E(\pi_i) = 1$.*

to a simple linear regression model with intercept $\beta_0^{SLR} = \beta_0^{\text{TRUE}} + E(\pi_i) = 0 + 2/b_\pi^{\text{TRUE}}$, slope coefficient $\beta_1^{\text{TRUE}} = 1$, but the error, $y_i - E(y_i \mid \mathbf{u}_i, \boldsymbol{\theta})$, though having mean zero and constant variance, $var(\pi_i) + \sigma_y^{2,\text{TRUE}} = 2/(b_\pi^{\text{TRUE}})^2 + 0.1^2$, is *not* normally distributed.

Figure 2 displays a boxplot of the M posterior expected values of the intercept under each analysis under scenario SLR: π-skewed with low variance, where we observe a large bias for estimation of the population model on the informative sample without including the sampling weights (Pop). Tables 1 and 2 present the bias, MSE, coverage of central 95% CIs and average lengths under both scenarios for each of the slope coefficient and the intercept, respectively. The figure and the left-hand columns 2 to 5 of the Table 1 show that when the true variance of π is relatively *low* the pseudo posterior yields a good point estimator of the slope coefficient with MSE lower even than under SRS. There is a trade-off of some bias for improved efficiency. The right-hand columns of the table, reporting the results under scenario SLR: π-skewed with relatively *high* variance, show that this property does not hold when the variance of π is large; the MSE for the pseudo posterior is now the highest, as is the bias, so there is no trade-off of one for the other. In other words, the point estimate of the fully Bayes model is robust against high variability of the inclusion probabilities while pseudo posterior estimator is not. In both scenarios, the average length of the CI is notably shorter under SRS than for the fully Bayesian approach under an IS because there is more variation in the realized samples drawn under an informative design, P_ν, than under SRS. Changing our focus to Table 2, the pseudo posterior underestimates the uncertainty (of the point estimate for the intercept); its CI fails to capture the nominal 95% coverage. In contrast, the fully Bayes model estimates the uncertainty appropriately; its CI maintains a coverage similar to the model under SRS. The population model without correction using the sampling weights (Pop) is unbiased for the slope coefficient, β_1 (but not the intercept), so it performs relatively well in Table 1. The Pop model expresses a huge bias, however, for the intercept, which is expected.

Table 1

Estimation of slope coefficient, $\beta_1^{TRUE} = 1$, under SLR: Bias, MSE, coverage of central 95% CI and its average length with three analyses under simulation scenarios SLR: π-skewed with low variance and with high variance.

True variance	Low ($b_\pi^{\text{TRUE}} = 2$)				High ($b_\pi^{\text{TRUE}} = 1$)			
of π_i	Full	Pseudo	Pop	SRS	Full	Pseudo	Pop	SRS
Bias	-0.006	0.020	-0.007	0.011	0.002	1.034	0.002	0.005
MSE	0.213	0.098	0.213	0.199	0.489	1.259	0.489	0.400
95% CI coverage	0.940	0.876	0.943	0.940	0.946	0.885	0.942	0.935
95% CI length	1.818	1.682	1.818	1.683	2.741	2.390	2.738	2.380

Table 2

Estimation of intercept, $\beta_0^{TRUE} + E(\pi) = 0 + 2/b_\pi^{TRUE}$, under SLR: Bias, MSE, coverage of central 95% CI and its average length with four analyses under simulation scenarios SLR: π-skewed with low variance and with high variance.

True variance	Low ($b_\pi^{\text{TRUE}} = 2$)				High ($b_\pi^{\text{TRUE}} = 1$)			
of π_i	Full	Pseudo	Pop	SRS	Full	Pseudo	Pop	SRS
Bias	-0.007	0.020	0.504	-0.005	-0.016	0.034	1.001	-0.002
MSE	0.317	0.368	1.077	0.312	0.164	0.191	1.157	0.137
95% CI coverage	0.953	0.876	0.522	0.941	0.962	0.892	0.285	0.928
95% CI length	1.123	0.966	1.048	0.971	1.710	1.371	1.577	1.373

All posterior sampling computations - for both the fully Bayes and pseudo posterior computations - were performed in Stan and there is no notable difference in the computation time between them because Stan does not leverage conjugacy. Were one to use a custom-built sampler, it would be likely that the computation per effective sample size would be lower for the pseudo posterior than the fully Bayes, though such would depend on the particular population model and the particular samplers.

4.1.2. Scenario SLR: π-symmetric.

We explore the robustness of our approach in the case where the inclusion probabilities are generated as symmetric but modeled as skewed. To do so, we repeat the simulation study in subsection 4.1.1, but now with $\pi_i \overset{\text{iid}}{\sim} \text{beta}(1.2, 1.2)$ in 1 (b). We call this simulation scenario: "SLR: π-symmetric". Notice that the fully Bayes approach, as in simulation scenario scenario SLR: π-skewed, misspecifies the distribution of $\pi_i \mid y_i, \boldsymbol{\kappa}$. As in scenario SLR: π-skewed, in subsection 4.1.1, the regression model for $y_i \mid \mathbf{u}_i, \boldsymbol{\theta}$ is misspecified; in particular the distribution of the error is *not* normally distributed.

Figure 3 and Table 3 are constructed in the same formats as Figure 2 and Table 2, for scenario: SLR: π-symmetric. The fully Bayesian and pseudolikelihood methods perform similarly in terms of bias, MSE but Pseudo credibitily

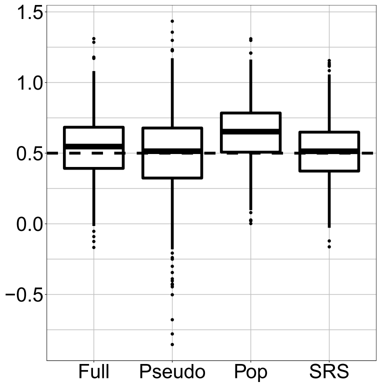

Fig 3. *Estimation of intercept regression coefficient under simulation scenario, SLR π-symmetric. Here, $\pi_i \overset{\text{iid}}{\sim}$ beta$(1.2, 1.2)$ in step 1(a) in Section 4.1.1. Box plots of posterior expected values of β_0 are displayed under IS for the fully Bayes, pseudo posterior approaches, ignoring the informative design (Pop), and under SRS. The horizontal dashed line represents simulation value of the intercept under the SLR model, i.e., $\beta_0^{TRUE} + E(\pi) = 0 + 1/2$.*

Table 3

Estimation of intercept, $\beta_0^{TRUE} + E(\pi) = 0 + 1/2$, under SLR: Bias, MSE, coverage of central 95% CI and its average length with four analyses under simulation scenario SLR: π-symmetric.

	Full	Pseudo	Pop	SRS
Bias	0.041	0.003	0.146	0.011
MSE	0.196	0.201	0.284	0.168
95% CI coverage	0.930	0.858	0.874	0.942
95% CI length	0.839	0.817	0.812	0.816

intervals undercover. Fully Bayes outperforms Pseudo, even though, by design, the population model for the inclusion probabilities is misspecified under the fully Bayes method because we have drawn the inclusion probabilities from a symmetric distribution and modeled them with a skewed distribution. The analysis ignoring IS, Pop, yields an estimator with higher bias and MSE. In both scenarios, SLR:π-skewed and SLR:π-symmetric, the fully Bayes CI mantains its nominal coverage and has average length wider than under SRS. That is, fully Bayes maintains its nominal coverage at the cost of wider credible intervals.

4.1.3. Scenario non-linear: π-skewed.

Our next simulation study explores the situation where the relationship between the predictor and the response is *not* linear. As before, the superindex TRUE indicates the simulation model parameters. To ease the computational burden, we simulate only one synthetic population dataset with $N = 10^5$ individuals. We run this simulation twice, first with sample size $n = 100$ and last with $n = 1000$. The objective is to estimate the curve $E(y \mid \mathbf{u})$ on a regular grid, $\mathbf{u} \in (0, 2)$. The analysis model is the splines regression model described in Subsection 2.3.

- Generate the population (once)
  1. Draw $\mathbf{u}_i \overset{\text{iid}}{\sim} \text{uniform}(0,2)$
  2. Draw $\pi_i \overset{\text{iid}}{\sim} \text{gamma}(2, rate = 1)$
  3. Generate the population response:

     $$y_i \mid \mathbf{u}_i, \pi_i, (\beta_{\mathbf{u}}^{\text{TRUE}}, \beta_{\pi}^{\text{TRUE}}, \beta_{\mathbf{u},\mathbf{u}}^{\text{TRUE}}, \sigma_y^{2,\text{TRUE}}) \sim \text{normal}(\beta_{\mathbf{u}}^{\text{TRUE}}\mathbf{u}_i + \beta_{\pi}^{\text{TRUE}}\pi_i + \beta_{\mathbf{u},\mathbf{u}}^{\text{TRUE}}\mathbf{u}_i^2, \sigma_y^{2,\text{TRUE}}),$$

     with $(\beta_{\mathbf{u}}^{\text{TRUE}}, \beta_{\pi}^{\text{TRUE}}, \beta_{\mathbf{u},\mathbf{u}}^{\text{TRUE}}, \sigma_y^{2,\text{TRUE}}) = (1, 1, -0.5, 0.1^2)$
- For $m = 1, \ldots, M$:
  1. Generate an IS and SRS of size n in the same fashion as in the SLR simulation study and, subsequently, estimate the model described in Section 2.3, using cubic B-splines, with the first and last internal knots equal to, respectively, the minimum and the maximum of the $\mathbf{u}_i$s in the sample.
  2. Compute the central 95% credible intervals for $E(y \mid \mathbf{u}) = B(\mathbf{u})E(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\pi})$, pointwise on a grid, *i.e.*, $\mathbf{u} = 0, 1/40, 2/40, \ldots, 80/40$, under the three analyses and store their credible interval lengths and an indicator for whether or not they contain the true curve given by

     $$\beta_{\mathbf{u}}^{\text{TRUE}}\mathbf{u} + \beta_{\pi}^{\text{TRUE}}E(\pi_i) + \beta_{\mathbf{u},\mathbf{u}}^{\text{TRUE}}\mathbf{u}^2 = \mathbf{u} + 2 - 0.5\mathbf{u}^2$$
- Based on the values stored in step 3 (a), estimate the (pointwise) Bias and MSE curves and based on the values stored in step 3 (b), estimate the coverage and average length of the central 95% CI for the curve $E(y \mid \mathbf{u})$ with $\mathbf{u} \in (0, 2)$.

Figure 4 shows the results of the simulation under this scenario with sample size, $n = 100$. The first plot panel (from the top) displays the simulation curve and those estimated under the three methods. The next four panels (from top-to-bottom) display the MSE, Bias, coverage of central 95% CI and the average length of the CI for the three methods, respectively. The SRS method performs uniformly the best in terms of fit, bias, MSE and coverage. It is followed by the fully Bayes approach. As in scenario SLR:π-skewed the pseudo posterior approach fails to maintain a 95% coverage for its central 95% CI. The bottom panel shows that the cost of having an IS is a wider CI for the same coverage due to the larger variation in information about the population in samples drawn under an informative design, as earlier discussed. The credibility intervals of the fully Bayes method achieve nominal coverage, even at the relatively low sample size of $n = 100$ with little-to-no bias, while the pseudo posterior expresses a relatively large amount of bias at $n = 100$. Figure 5 is the same as Figure 4 after increasing the sample size to $n = 1000$. The results are qualitatively the same as when $n = 100$ except that the pseudo posterior performs better in terms of bias; in agreement with the result that the pseudo posterior approach produces asymptotically unbiased estimators (Savitsky and Toth, 2016). Even at $n = 1000$, the pseudo posterior continues to under-cover.

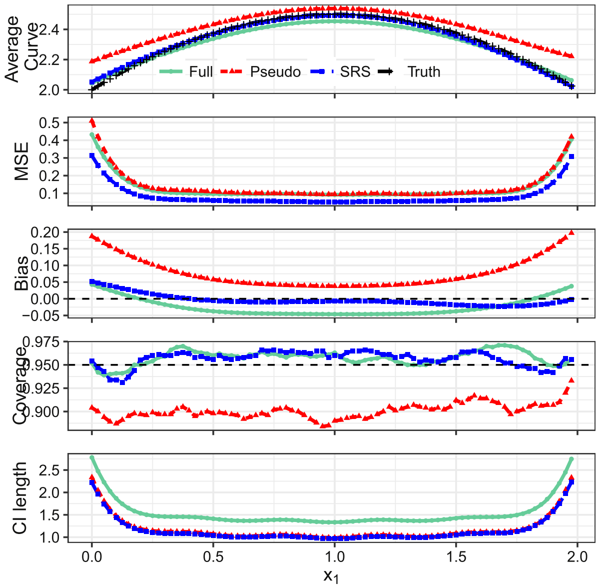

Fig 4. *Simulation scenario non-linear: π-skewed with sample size $n = 100$. From top to bottom, average estimated curve, MSE, bias, coverage probability of 95% central credible intervals and their average length. The black curve in the upper panel is the true curve. For visual purposes the horizontal (discontinuous) lines at 0 and 0.95 in third and fourth panels, respectively, are depicted.*

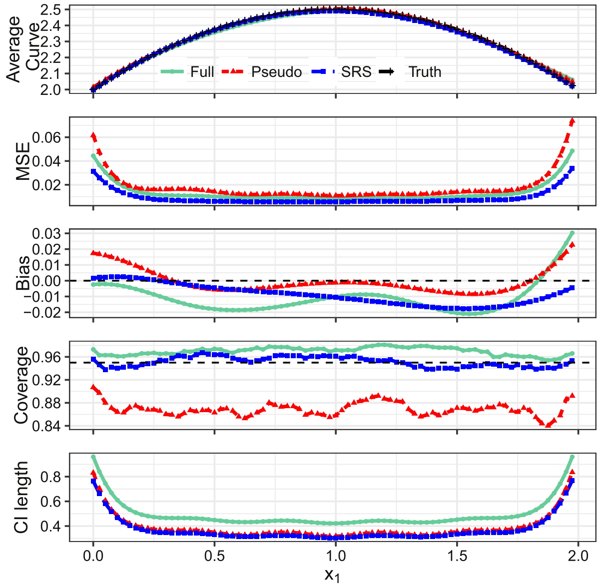

Fig 5. *Simulation scenario non-linear:π-skewed with sample size $n = 1000$. Same as Figure 4 but with $n = 1000$.*

4.2. Robustness of population model for $\pi|y$

In Section (4.1.2) we generate the conditional distribution for $\pi|y$ from a symmetric beta distribution, rather than the skewed lognormal distribution that we assume for the population generating model. Our intent is to examine the robustness of the fully Bayes method to misspecification of the population generating model for $\pi|y$. This set-up demonstrates excellent robustness for the coverage of the credibility intervals under a simple misspecification of the population model for $\pi|y$. We next conduct a further assessment of robustness of lognormal population model specification for $\pi|y$ under a real world condition where we don't know the population generating model for $\pi|y$. Most sampling designs are multistage. While a distribution for the inclusion probabilities may be specified at each stage (e.g., based on a proportion-to-size design for the inclusion probabilities in each stage), the resulting induced population distribution of the $\pi|y$ may not conform to a simple lognormal due to the algorithmic construction of the sampling design. We construct a test of robustness of the fully Bayes method under this set-up by treating our 6847 sample of individuals from the 2013-2014 National Health and Nutrition Examination Survey (NHANES) (which we more fully explain in Section (5)) as a *population*. So $N = 6847$. The NHANES sampling design is multistage with the sampling of geographically-indexed primary sampling units under unequal inclusion probabilities. We proceed to draw samples from this putative population using the inverse of published unit sampling weights as inclusion probabilities. Our variable of interest is the log transformed systolic blood pressure (SBP) measurements of participants. Here, we focus on estimation of $\beta_0^{\text{TRUE}} = (1/N)\sum_{i=1}^{N} y_i = 4.762395$, the putative population mean. We use a particular case of the linear regression model in Subsection (2.1) for our analysis. Specifically, the population model is constructed as, $y_i \mid \boldsymbol{\theta} \sim \text{normal}(\beta_0, \sigma_y^2)$ with $\boldsymbol{\theta} = (\beta_0, \sigma_y^2)$. The full estimation model assumes $\pi_i \mid y_i, \boldsymbol{\kappa} \sim \text{lognormal}(\kappa_0 + \kappa_y y_i, \sigma_\pi^2)$, this is $\boldsymbol{\kappa} = (\kappa_0, \kappa_y, \sigma_\pi^2)$, which is misspecified. As in Subsection 4.1, we compare the estimators yielded by the fully Bayesian, pseudoposterior and ignoring IS (Pop) using SRS as gold standard.

- For $m = 1, \ldots, M = 10^3$, Monte Carlo iterations
  1. Draw two samples with replacement of size n:
     (a) Take an IS with, $\Pr[(y_i, \pi_i) \in \text{sample}] = \pi_i / \sum_{i'=1}^{N} \pi_{i'}$.
     (b) Take an SRS sample.
  2. Conduct estimation using the priors given in Equation (8).
     (a) The informative sample is estimated using,
        i. Full: the fully Bayes approach.
        ii. Pseudo: the pseudo posterior formulation.
        iii. Pop: the population model (ignoring IS).
     (b) SRS: The SRS is estimated using the population model.
  3. (a) Store the posterior expected values of β_0 under the analyses 2(a) i-iii and 2(b).

   (b) Also compute the central 95% credible interval (CI) for β_0 under the four analyses and store their lengths and indicator of whether they contain β_0^{TRUE}.

- Using the values stored in step 3, (a) estimate the Bias and MSE of the estimate of β_0^{TRUE} and, (b), estimate the coverage and average length of the central 95% credible interval (CI) for β_0^{TRUE}.

Table 4 compares bias, MSE and coverage properties among the fully Bayes estimator, the pseudo posterior (Pseudo), estimation of the population model on the informative sampling without weighting (Pop) and a simple random sample (SRS) taken from the same population. We employ an increasing sequence of low sample sizes, $(20, 50, 100)$, to examine convergence properties. The results in Table 4, as before, demonstrate that the fully Bayes (Full) model achieves nominal frequentist coverage of the credibility interval at the expense of a longer interval, while the credibility intervals for the pseudo posterior (Pseudo) and estimation of the populating generating model without correcting for informative sampling both under cover. The improvement rate in MSE contracts somewhat faster for the fully Bayes than it does for the pseudo posterior, which would be predicted by our theory since the convergence rate is injured by $\gamma = \max_i \pi_i$, which is a smoothed (modeled) quantity in the fully Bayes case and an unmodeled, raw, quantity for the pseudo posterior.

Table 4

Simulation study result using the $N = 6874$ NHANES observations as a population from which are drawn samples of size, $n = 20, 50$ and 100. For visual purposes the MSE$\times 10^3$, as opposed to the MSE, is reported.

	Full	Pseudo	Pop	SRS
		$n = 20$		
Bias	0.012	0.000	-0.023	-0.000
MSE $\times 10^3$	2.170	1.905	1.586	0.972
95% CI coverage	0.953	0.851	0.870	0.958
95% CI length	0.199	0.131	0.135	0.136
		$n = 50$		
Bias	0.007	-0.001	-0.023	0.000
MSE $\times 10^3$	0.792	0.795	0.990	0.385
95% CI coverage	0.942	0.841	0.770	0.957
95% CI length CI	0.109	0.080	0.082	0.082
		$n = 100$		
Bias	0.007	-0.000	-0.024	0.000
MSE $\times 10^3$	0.390	0.406	0.762	0.218
95% CI coverage	0.939	0.845	0.631	0.946
95% CI length	0.075	0.056	0.057	0.057

5. Application

As an illustration we explore the relationship between caffeine and systolic blood pressure (SBP), applying our methodology to data from the 2013-2014 National Health and Nutrition Examination Survey (NHANES). We will perform two analyses. In the first analysis we do not adjust for any other covariate and in the second we adjust for age and gender.

Due in part to technical, budget and logistic considerations, the survey design of the NHANES is complex. NHANES is designed to assess the health and nutritional status of adults and children in the United States (CDC-A, 2016). Data are collected from a sample of individuals from the non-institutionalized, civilian US population. Although nationally representative, the NHANES is designed to oversample specific subpopulations of interest (e.g. minorities, low income groups, pregnant women) for population-based studies using a complex, multistage, cluster sampling design (CDC-B, 2016). To correct for informativeness (partly induced by the correlation between response values and group memberships), the NHANES survey data are released with observation-indexed sampling weights (based on marginal inclusion probabilities) corresponding to the sampled participants. These sampling weights are computed based on the sampling design (i.e. probability of the individual of being included in the sample), non-response and the masking of the individual for confidentiality (CDC-D, 2016).

NHANES participants are required to visit a mobile examination center where Health and nutrition information are collected during these visits. In particular, SBP is measured for participants of ages 8 years and older. A 24-hour dietary recall interview is taken. The results of this interview are used to estimate the participant's intakes of nutrients during the 24-hour period preceding their interview (from the period of midnight to midnight). Our dataset consists of systolic blood pressure (SBP) (mm Hg) measurements of $n = 6847$ participants 8 or older and 24-hour caffeine consumption (mg), along with their sampling weights, as estimated from the first 24-hour dietary recall interview. See Appendix D for details regarding our data preprocessing.

We fit the spline basis model introduced in Subsection 2.3 with response, $y = \log(\text{SBP})$, inclusion probability, π, designed to be proportional to the inverse of the NHANES sampling weight, and predictor, $x = \log(\text{caffeine consumption}+1)$. We fit both the fully Bayesian (Full) and the pseudo posterior (Pseudo) models. Additionally, against the recommendation stated in NHANES, we fit the splines model for the population that ignores IS (label "Pop", equivalent to the pseudo posterior model with all sampling weights equal to 1). The cubic B-spline model is constructed with the priors and hyperparameters described in Section 2.3.

After a burn-in period of 10^4 iterations, a Monte Carlo posterior sample of size 10^4 was retained for each of the model parameters and used to estimate $E(y \mid x)$, together with 95% credible intervals on a pointwise basis under the three models. Figure 6 displays these curves. There is a positive relationship between caffeine consumption and blood pressure that levels off in higher ranges of caffeine consumption. The Pseudo posterior estimates a gradually increasing relationship, while the fully Bayes estimates 3 distinct regions in the support of log(caffeine+1), where each expresses a different sensitivity to log(SBP); in particular, the credible intervals for these two models do not overlap in parts of the second region (from left-to-right) in a range of $2.5 \leq x \leq 4$, where the

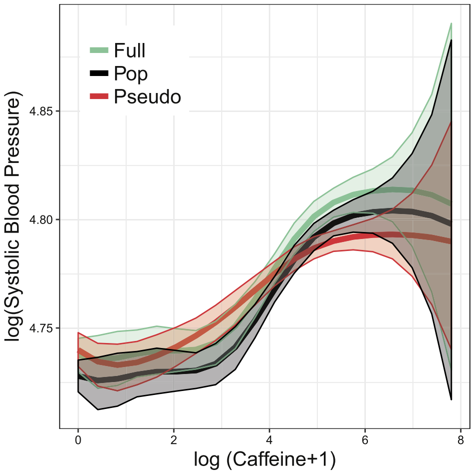

Fig 6. *Caffeine consumption (mg) v.s. SBP (mm Hg) under the Full, Pseudo and Pop (Ignoring IS) models.*

fully Bayes model shows a notably greater sensitivity. In this case, then, the inferences performed under these two models would differ.

The resulting negative sign of κ_y, with central 95% credible interval (-0.64,-0.35) and $\Pr(\kappa_y < 0 \mid \text{data}) \approx 1$, in the model for $\pi_i \mid y_i, x_i$ indicates that the sample design is informative for the logarithm of SBP (when not conditioned on predictors). The higher the value of the response the lower is the inclusion probability.

We next proceed to check if this positive association between SBP and caffeine consumption still holds after controlling for age and gender. We categorized the age into seven groups as shown in Table 5. The reference categories were the 8-17 age group and the male gender. We adjust the cubic B-spline model

$$y_i \sim \text{normal}\left(\mu(\mathbf{u}_i) + \beta^{age}_{age(i)} + \beta^{gender}_{gender(i)}, \sigma^2_y\right)$$

with $\mu(\mathbf{u}_i)$ defined in Equation (10). $\beta^{age}_{age(i)}$ and $\beta^{gender}_{gender(i)}$ are the regression coefficients associated with the age group and gender of the participant i for groupings, $age(i) \in \{1, \ldots, 7\}$ and $gender(i) \in \{1, 2\}$. The regression coefficients associated to the reference groups are set to 0 ($\beta^{age}_1 = \beta^{gender}_1 \equiv 0$). We employ vague priors $\beta^{age}_2, \ldots, \beta^{age}_7, \beta^{gender}_2 \overset{\text{iid}}{\sim} \text{normal}(0, 10^4)$. Age group and gender were also included as covariates in $\pi_i \mid y_i, \mathbf{v}_i$, where $\mathbf{v}_i$ is a vector of dimension 9, by incorporating for an intercept, $x \equiv \log(\text{caffeine} + 1)$, seven age groups and gender. The estimated curves along with their 95% credible intervals for the reference group (males between 8 and 16 years old) under the three models are shown in Figure 7. The positive association between caffeine consumption and SBP vanishes when controlling for age and gender. The estimated curves under the fully Bayes and Pseudo posterior models differ in shape but

Table 5

Fully Bayes model parameter estimates for caffeine consumption versus SBP. (Caffeine) spline parameters $\beta_2, \dots, \beta_8$ and σ_β are not depicted to save space.

	mean	**sd**	**2.5%**	**97.5%**
Parameters of $y_i \mid \cdots$				
Regression coefficients				
Gender (female)	-0.03	0.00	-0.04	-0.03
Age 17-24	0.08	0.00	0.07	0.09
Age 25-34	0.09	0.01	0.08	0.10
Age 35-39	0.12	0.01	0.11	0.13
Age 40-49	0.14	0.01	0.13	0.15
Age 50-59	0.18	0.01	0.17	0.19
Age 60 or older	0.24	0.00	0.23	0.25
σ_y	0.11	0.00	0.11	0.12
(Caffeine) Spline parameters in $y_i \mid \cdots$				
β_1	4.66	0.03	4.59	4.73
$\vdots$				
Parameters of $\pi_i \mid y_i, \cdots$				
Intercept	-0.91	0.41	-1.71	-0.11
κ_y	0.11	0.09	-0.06	0.28
Caffeine	-0.05	0.01	-0.06	-0.04
Gender (Female)	-0.03	0.02	-0.07	0.01
Age 17-24	-0.55	0.04	-0.62	-0.48
Age 25-34	-0.74	0.04	-0.81	-0.67
Age 35-39	-0.67	0.05	-0.77	-0.58
Age 40-49	-0.67	0.04	-0.75	-0.59
Age 50-59	-0.72	0.04	-0.80	-0.63
Age 60 or older	-0.54	0.04	-0.62	-0.47
σ_π	0.83	0.01	0.81	0.84

their bands overlap across all the range of x. The posterior mean curve for the fully Bayes model is, however, more smoothly centered on the horizontal line to more strongly indicate little-to-no association between systolic blood pressure and caffeine consumption for the reference group. The credible intervals for the fully Bayes model and Pop (ignoring IS) are almost perfectly overlapping, which suggests a non-informative sampling design. The central 95% credible interval for κ_y, $(-.06, 0.28)$, contains zero, confirming that the sampling design is non-informative for SBP when controlling for age and gender. Table 5 depicts the posterior mean, standard deviation and 2.5% and 97.5% quantiles of the marginal posterior distribution of the fully Bayes, adjusting for age and gender, model parameters. The regression coefficients in the conditional model for $y_i \mid \mathbf{u}_i$ indicate that mean SBP for females is lower and SBP increases with age.

When the sampling design is not informative for the chosen response (given the available predictors), the fully Bayes model provides similar inference than under SRS. By contrast, the pseudo posterior estimator is notably noisier than under SRS, which in some cases may lead to incorrect inference.

Steffen et al. (2012) performed a meta-analysis of randomized clinical trials and prospective studies and concluded that coffee consumption was not associated with a significant change in SBP. The fully Bayes and Pop curves in Figure 7 are nearly horizontal, which effectively reproduces this result. The pseudo

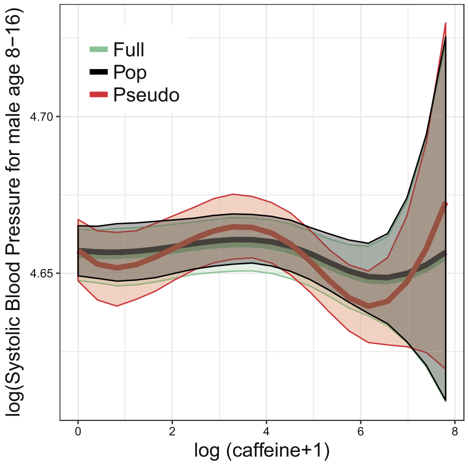

FIG 7. *Caffeine consumption (mg) v.s. SBP (mm Hg), for 8-16 years old males, under the Full, Pseudo and Pop (Ignoring IS) models.*

posterior result, however, expresses a lot more oscillation around around the horizontal line.

6. Discussion

We have developed a novel fully Bayes approach to incorporate the sampling weights into the estimation of population model parameters on data collected under an informative sampling design by specifying a joint population distribution for the response of interest and sampling inclusion probabilities. The approach uses only quantities observed in sample, and discards variation in weights not dependent on response. We have shown via simulation that the method performs as well and often better in terms of bias, MSE and, the coverage of its central 95% credible interval than the non-fully Bayesian plug-in pseudo posterior. Our fully Bayes approach demonstrates a consistently superior ability to accurately measure uncertainty, in contrast with the pseudo posterior method, which is too confident such that is fails to achieve nominal coverage. The price to be paid for the achievement of unbiased inference with correct coverage of population model parameters from an informative sample is wider credible intervals to achieve the same coverage relative to simple random sampling because there is more variation in the information content around the population in informative samples. Now that we have derived and established the utility of the fully Bayesian method, we plan to examine multi-stage sampling designs that include a cluster step (which induces dependence among sampled units under P_ν) in order to explore the coverage properties of the credibility intervals for our fully Bayesian estimators. We will parameter random effects in our population model for y in order to capture the within cluster dependence.

We applied our fully Bayes approach to explore the relationship between caffeine consumption and SBP analyzing NHANES data. We notice a (non linear) positive association between this two variables. This association vanishes when controlling for age and gender. The fully Bayesian approach estimates a smoother and closer to the horizontal line (indicating no association) caffeine *v.s.* SBP curve as compared to the pseudo posterior approach, which better agrees with previous research for the relationship between caffeine consumption and SBP.

The unusual assumption of our approach, from the perspective of the survey sampler, is that it considers the weights (inclusion probabilities) as random (with respect to P_0). Although the fully Bayes approach requires specification of a conditional population distribution for the inclusion probabilities, we have shown that coverage properties are robust to misspecification of this distribution. A disadvantage of the proposed method is that it requires a customized posterior sampler because the posterior sampler designed for simple random sampling cannot easily be adapted to IS under the fully Bayes construction. We cope with this problem by relying on Stan (Carpenter et al., 2016). A related issue is the required integration step for every observation at every iteration of the posterior sampler. This computation must be performed, numerically, in the general case. Theorem 1 provides some conditions in the sampling distribution and the conditional distribution of the weights given the response that allow a closed form for the fully Bayes likelihood. We applied the theorem to illustrate a collection of useful population joint distributions over the inclusion probabilities and response that all allow for a closed form for the integration step. More research is needed to apply our approach to more complicated settings. Future work will focus on performing the computation of this expected value, numerically, or circumventing it by introducing latent random variables.

Appendix A: Proof of Theorem 1

Recall that the vector of covariates $\mathbf{x}_i$ has been split into two sets of covariates, $\mathbf{u}_i$ and $\mathbf{v}_i$, relevant for the distribution of $y_i \mid \boldsymbol{\theta}, \mathbf{u}_i$, and $\pi_i \mid y_i, \mathbf{v}_i, \boldsymbol{\kappa}$, respectively. Further recall the definition for $M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}) := E_{y|\mathbf{u}_i,\boldsymbol{\theta}}\left[\exp\left\{g(y_i, \mathbf{v}_i, \boldsymbol{\kappa})\right\}\right]$. Using the fact that $v \sim \text{lognormal}(m, v^2) \Rightarrow E(v) = \exp(m + v^2/2)$, we obtain,

$$\begin{aligned}
E_{y|\mathbf{u}_i,\boldsymbol{\theta}}\left\{E\left[\pi_i \mid y, \mathbf{v}_i, \boldsymbol{\kappa}\right]\right\} =& E_{y|\mathbf{u}_i,\boldsymbol{\theta}}\left[\exp\{g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) + t(\mathbf{v}_i, \boldsymbol{\kappa}) + \sigma_\pi^2/2\}\right] \\
=& \exp\left[t(\mathbf{v}_i, \boldsymbol{\kappa}) + \sigma_\pi^2/2\right] \times E_{y|\mathbf{u}_i,\boldsymbol{\theta}}\left[\exp\{g(y_i, \mathbf{v}_i, \boldsymbol{\kappa})\}\right] \\
=& \exp\left[t(\mathbf{v}_i, \boldsymbol{\kappa}) + \sigma_\pi^2/2\right] \times M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta})
\end{aligned}$$

Noting that $\text{lognormal}(\pi \mid m, v^2) = \text{normal}(\log \pi \mid m, v^2) \times 1/\pi$, plugging the expression above in equation (5) we obtain,

$$p_s(y_i, \pi_i \mid \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}, \boldsymbol{\kappa}) = \frac{\pi_i p(\pi_i \mid y_i, \mathbf{v}_i, \boldsymbol{\kappa})}{E_{y|\mathbf{u}_i,\boldsymbol{\theta}}\left\{E\left[\pi_i' \mid y, \mathbf{v}_i, \boldsymbol{\kappa}\right]\right\}} p(y_i \mid \mathbf{u}_i, \boldsymbol{\theta})$$

$$= \frac{\pi_i \times \text{lognormal}\left(\pi_i \mid t(\mathbf{v},\boldsymbol{\kappa}) + g(y_i,\mathbf{v},\boldsymbol{\kappa}), \sigma_\pi^2\right)}{\exp\left[t(\mathbf{v}_i,\boldsymbol{\kappa}) + \sigma_\pi^2/2\right] \times M_y(\boldsymbol{\kappa};\mathbf{u}_i,\mathbf{v}_i,\boldsymbol{\theta})} \times p(y_i \mid \mathbf{u}_i,\boldsymbol{\theta})$$

$$= \frac{\cancel{\pi}_i \times (1/\cancel{\pi}_i) \times \text{normal}(\log \pi_i \mid t(\mathbf{v},\boldsymbol{\kappa}) + g(y_i,\mathbf{v},\boldsymbol{\kappa}), \sigma_\pi^2)}{\exp\left[t(\mathbf{v}_i,\boldsymbol{\kappa}) + \sigma_\pi^2/2\right] \times M_y(\boldsymbol{\kappa};\mathbf{u}_i,\mathbf{v}_i,\boldsymbol{\theta})} \times p(y_i \mid \mathbf{u}_i,\boldsymbol{\theta}),$$

proving the theorem.

Appendix B: Closed-Form for log likelihood of simple linear regression model

$$\begin{aligned}\log p_s(y_i,\pi_i \mid \mathbf{u}_i,\mathbf{v}_i,\boldsymbol{\theta},\boldsymbol{\kappa}) \propto & -\frac{1}{2}\log(\sigma_\pi^2) - \frac{1}{2\sigma_\pi^2}\left[\log \pi_i - \left(\kappa_y y_i + \mathbf{v}_i^t\boldsymbol{\kappa}_x\right)\right]^2 \\ & -\frac{1}{2}\log(\sigma_y^2) - \frac{1}{2\sigma_y^2}\left[y_i - \mathbf{u}_i^t\boldsymbol{\beta}\right]^2 \\ & -\mathbf{v}_i^t\boldsymbol{\kappa}_x - \sigma_\pi^2/2 - \kappa_y\mathbf{u}_i^t\boldsymbol{\beta} - \kappa_y^2\sigma_y^2/2\end{aligned}$$

$$\begin{aligned}\log p_s(\mathbf{y},\boldsymbol{\pi} \mid \mathbf{x},\boldsymbol{\theta},\boldsymbol{\kappa}) \propto & -\frac{n}{2}\log(\sigma_\pi^2) - \frac{1}{2\sigma_\pi^2}\sum_i\left[\log \pi_i - \left(\kappa_y y_i + \mathbf{v}_i^t\boldsymbol{\kappa}_x\right)\right]^2 \\ & -\frac{n}{2}\log(\sigma_y^2) - \frac{1}{2\sigma_y^2}\sum_i\left[y_i - \mathbf{u}_i^t\boldsymbol{\beta}\right]^2 \\ & -\sum_i\mathbf{v}_i^t\boldsymbol{\kappa}_x - n\sigma_\pi^2/2 - \kappa_y\sum_i\mathbf{u}_i^t\boldsymbol{\beta} - n\kappa_y^2\sigma_y^2/2\end{aligned}$$

Appendix C: Enabling Lemmas for Theorem 2

We next construct two enabling results needed to prove Theorem 2 to account informative sampling under (A4), (A5) and (A6). The first enabling result is used to bound from above the numerator in the expression for the expectation with respect to the joint distribution for population generation and the taking of the informative sample, (P_{λ_0}, P_ν), of the fully Bayesian, sampling-weighted posterior distribution in Equation (17) on the restricted set of measures that includes those P_λ that are at some *minimum* distance, $\delta\xi_{N_\nu}$, from P_{λ_0} under pseudo Hellinger metric, $d^\pi_{N_\nu}$. The second result, Lemma 4, extends Lemma 8.1 of Ghosal et al. (2000) to bound the probability of the denominator of Equation (17) with respect to (P_{λ_0}, P_ν), from below. Given these 2 Lemmas, the subsequent proof exposition for Theorem 2 is identical to Savitsky and Toth (2016).

Lemma 3. *Suppose conditions (A1) and (A4) hold. Then for every $\xi > \xi_{N_\nu}$, a constant, $K > 0$, and any constant, $\delta > 0$,*

$$\mathbb{E}_{P_0,P_\nu}\left[\int_{P\in\mathcal{P}\backslash\mathcal{P}_{N_\nu}}\prod_{i=1}^{N_\nu}\frac{p^\pi}{p_0^\pi}\left(x_{\nu i}\delta_{\nu i}\right)d\Pi\left(\lambda\right)\left(1-\phi_{n_\nu}\right)\right] \leq \gamma\Pi\left(\Lambda\backslash\Lambda_{N_\nu}\right) \qquad (20)$$

$$\mathbb{E}_{P_{\lambda 0},P_{\nu}}\left[\int\limits_{\lambda\in\Lambda_{N_{\nu}}:d^{\pi}_{N_{\nu}}\left(P_{\lambda},P_{\lambda_0}\right)>\delta\xi}\prod_{i=1}^{N_{\nu}}\frac{p^{\pi}}{p^{\pi}_{0}}\left(x_{\nu i}\delta_{\nu i}\right)d\Pi\left(\theta\right)\left(1-\phi_{n_{\nu}}\right)\right]\leq$$

$$2\gamma^{2}\exp\left(\frac{-Kn_{\nu}\delta^{2}\xi^{2}}{\gamma}\right). \qquad (21)$$

This result is adjusted from that in Savitsky and Toth (2016) by multiplying the upper bounds on the right-hand side of both equations by constant multiplier, $\gamma \geq 1$, defined in condition (A4).

Proof. The proof approach is the same as Savitsky and Toth (2016), where we first bound the left-hand sides of Equations (20) and (21) from above by an expectation of the test statistic, ϕ_{n_ν}. When then further refine this bound on the the two subsets of measures outlined in those equations. The refinement step is unchanged from Savitsky and Toth (2016), but the first step is revised due to our unique form for our sampling-weighted, fully Bayesian estimator, outlined in Equation (16). So we fully repeat the first step in the proof, below.

Fixing ν, we index units that comprise the population with, $U_\nu = \{1,\ldots,N_\nu\}$. Next, draw a single observed sample of n_ν units from U_ν, indexed by subsequence,
$\{i_\ell \in U_\nu : \delta_{\nu i_\ell} = 1,\ \ell = 1,\ldots,n_\nu\}$. Without loss of generality, we simplify notation to follow by indexing the observed sample, sequentially, with $\ell = 1,\ldots,n_\nu$.

We next decompose the expectation under the joint distribution with respect to population generation, P_{λ_0}, and the drawing of a sample, P_ν,

Suppose we draw λ from some set $B \subset \Lambda$. By Fubini,

$$\mathbb{E}_{P_{\lambda_0},P_{\nu}}\left[\int\limits_{P\in B}\prod_{i=1}^{N_{\nu}}\frac{p^{\pi}_{\lambda}}{p^{\pi}_{\lambda_0}}\left(x_{\nu i}\delta_{\nu i}\right)d\Pi\left(\lambda\right)\left(1-\phi_{n_{\nu}}\right)\right]$$

$$\leq\int\limits_{\lambda\in B}\left[\mathbb{E}_{P_0,P_{\nu}}\prod_{i=1}^{N_{\nu}}\frac{p^{\pi}_{\lambda}}{p^{\pi}_{\lambda_0}}\left(x_{\nu i}\delta_{\nu i}\right)\left(1-\phi_{n_{\nu}}\right)\right]d\Pi\left(\lambda\right) \qquad (22)$$

$$\leq\int\limits_{\lambda\in B}\left\{\sum_{\boldsymbol{\delta}_{\nu}\in\Delta_{\nu}}\mathbb{E}_{P_{\lambda_0}}\left[\prod_{\ell=1}^{n_{\nu}}\left[\frac{\pi_{\nu\ell}}{\pi^{\lambda}_{\nu\ell}}\frac{p_{\lambda}}{p_{\lambda_0}}\left(x_{\nu\ell}\right)\right]\left(1-\phi_{n_{\nu}}\right)\middle|\boldsymbol{\delta}_{\nu}\right]P_{P_{\nu}}\left(\boldsymbol{\delta}_{\nu}\right)\right\}d\Pi\left(\lambda\right) \qquad (23)$$

$$\leq\int\limits_{\lambda\in B}\left\{\sum_{\boldsymbol{\delta}_{\nu}\in\Delta_{\nu}}\mathbb{E}_{P_{\lambda_0}}\left[\prod_{\ell=1}^{n_{\nu}}\left[\frac{1}{\pi^{\lambda}_{\nu\ell}}\frac{p_{\lambda}}{p_{\lambda_0}}\left(x_{\nu\ell}\right)\right]\left(1-\phi_{n_{\nu}}\right)\middle|\boldsymbol{\delta}_{\nu}\right]P_{P_{\nu}}\left(\boldsymbol{\delta}_{\nu}\right)\right\}d\Pi\left(\lambda\right) \qquad (24)$$

$$\leq\int\limits_{\lambda\in B}\max_{\boldsymbol{\delta}_{\nu}\in\Delta_{\nu}}\mathbb{E}_{P_{\lambda_0}}\left[\prod_{\ell=1}^{n_{\nu}}\left[\frac{1}{\pi^{\lambda}_{\nu\ell}}\frac{p_{\lambda}}{p_{\lambda_0}}\left(x_{\nu\ell}\right)\right]\left(1-\phi_{n_{\nu}}\right)\middle|\boldsymbol{\delta}_{\nu}\right]d\Pi\left(\lambda\right) \qquad (25)$$

$$\leq\int\limits_{\lambda\in B}\mathbb{E}_{P_{\lambda_0}}\left[\prod_{\ell=1}^{n_{\nu}}\left[\frac{1}{\pi^{\lambda}_{\nu\ell}}\frac{p_{\lambda}}{p_{\lambda_0}}\left(x_{\nu\ell}\right)\right]\left(1-\phi_{n_{\nu}}\right)\middle|\boldsymbol{\delta}^{*}_{\nu}\right]d\Pi\left(\lambda\right) \qquad (26)$$

$$\leq \gamma \int\limits_{\lambda \in B} \mathbb{E}_{P_{\lambda_0}} \left[\prod_{\ell=1}^{n_\nu} \left[\frac{p_\lambda}{p_{\lambda_0}} (x_{\nu\ell}) \right] (1 - \phi_{n_\nu}) \middle| \boldsymbol{\delta}_\nu^* \right] d\Pi(\lambda) \tag{27}$$

$$\leq \gamma \int\limits_{\lambda \in B} \mathbb{E}_{\boldsymbol{\delta}_\nu^*} (1 - \phi_{n_\nu}) \, d\Pi(\lambda),$$

where $\sum_{\boldsymbol{\delta}_\nu \in \Delta_\nu} P_{P_\nu}(\boldsymbol{\delta}_\nu) = 1$ (Särndal et al., 2003) and $\boldsymbol{\delta}_\nu^* \in \Delta_\nu = \left\{ \{\delta_{\nu i}^*\}_{i=1,\ldots,N_\nu}, \ \delta_{\nu i}^* \in \{0,1\} \right\}$ denotes that sample, drawn from the space of all possible samples, Δ_ν, which maximizes the probability under the population generating distribution for the event of interest. The inequality in Equation (27) results from the bound, $\frac{1}{\pi_{\nu\ell}^\lambda} \leq \gamma$, specified in Condition (A4).

The remainder (second step) of the proof is the same as in Savitsky and Toth (2016), except one multiplies their result by γ to compute the bounds for Equations (20) and (21). ☐

Lemma 4. *For every $\xi > 0$ and measure Π on the set,*

$$B = \left\{ \lambda : -\mathbb{E}_{\lambda_0} \log \left(\frac{p_\lambda}{p_{\lambda_0}} \right) \leq \xi^2, \mathbb{E}_{\lambda_0} \left(\log \frac{p_\lambda}{p_{\lambda_0}} \right)^2 \leq \xi^2 \right\}$$

under the conditions (A2), (A3), (A4), and (A5), we have for every $C > 0$ and N_ν sufficiently large,

$$Pr \left\{ \int\limits_{\lambda \in \Lambda} \prod_{i=1}^{N_\nu} \frac{p_\lambda^\pi}{p_{\lambda_0}^\pi} (x_{\nu i} \delta_{\nu i}) \, d\Pi(\lambda) \leq \exp\left[-(1+C) N_\nu \xi^2 \right] \right\} \leq \frac{\gamma + C_3}{C^2 N_\nu \xi^2}, \tag{28}$$

where the above probability is taken with the respect to P_{λ_0} and the sampling generating distribution, P_ν, jointly.

The sum of positive constants, $\gamma + C_3$, is greater than 1 and will be larger for sampling designs where the (modeled) inclusion probabilities, $\{\pi_{\nu i}^\lambda\}$, are more variable.

Proof. The first part of the proof bounds the integral on the left-hand side of the event over which we take the probability (in Equation (28)), from below, by a centered and scaled empirical process. This first part is altered under our construction for our sampling-weighted, fully Bayesian estimator, outlined in Equation (16). So we specify this part, here.

By Jensen's inequality,

$$\log \int\limits_{\lambda \in \Lambda} \prod_{i=1}^{N_\nu} \frac{p_\lambda^\pi}{p_{\lambda_0}^\pi} (x_{\nu i} \delta_{\nu i}) \, d\Pi(\lambda) \geq \sum_{i=1}^{N_\nu} \int\limits_{\lambda \in \Lambda} \log \frac{p_\lambda^\pi}{p_{\lambda_0}^\pi} (x_{\nu i} \delta_{\nu i}) \, d\Pi(\lambda)$$

$$= N_\nu \cdot \mathbb{P}_{N_\nu} \int\limits_{\lambda \in \Lambda} \log \frac{p_\lambda^\pi}{p_{\lambda_0}^\pi} d\Pi(\lambda),$$

where we recall that the last equation denotes the empirical expectation functional taken with respect to the joint distribution over population generating and informative sampling. By Fubini,

$$\begin{aligned}
\mathbb{P}_{N_\nu} \int_{\lambda\in\Lambda} \log \frac{p_\lambda^\pi}{p_{\lambda_0}^\pi} d\Pi(\lambda) &= \int_{\lambda\in\Lambda} \left[\mathbb{P}_{N_\nu} \log \frac{p_\lambda^\pi}{p_{\lambda_0}^\pi}\right] d\Pi(\lambda) \qquad (29)\\
&= \int_{\lambda\in\Lambda} \left[\mathbb{P}_{N_\nu} \delta_\nu \log \left\{\frac{\pi_\nu}{\pi_\nu^\lambda} \frac{p_\lambda}{p_{\lambda_0}}\right\}\right] d\Pi(\lambda) \qquad (30)\\
&\geq \int_{\lambda\in\Lambda} \left[\mathbb{P}_{N_\nu} \frac{\delta_\nu}{\pi_\nu^\lambda} \log \frac{p_\lambda}{p_{\lambda_0}}\right] d\Pi(\lambda) \qquad (31)\\
&= \int_{\lambda\in\Lambda} \left[\mathbb{P}_{N_\nu}^\pi \log \frac{p_\lambda}{p_{\lambda_0}}\right] d\Pi(\lambda) \qquad (32)\\
&= \mathbb{P}_{N_\nu}^\pi \int_{\lambda\in\Lambda} \log \frac{p_\lambda}{p_{\lambda_0}} d\Pi(\lambda), \qquad (33)
\end{aligned}$$

where we, again, apply Fubini. Equation (30) provides an upper bound for Equation (31) because $\pi_{\nu i} \leq 1$, $\frac{1}{\pi_{\nu i}^\lambda} > 0$, while $\log \frac{p_\lambda^\pi}{p_{\lambda_0}^\pi}(x_{\nu i}) < 0$.

The remainder of the proof, which uses Chebyshev to provide an upper bound on the probability of the event in Equation (28), is specified by,

$$\leq \xi^2 \sup_\nu \left[\frac{1}{\min\limits_{i\in U_\nu} \pi_{\nu i}^\lambda}\right] + \xi^2 (N_\nu - 1) \sup_\nu \max_{i\neq j\in U_\nu} \left[\left|\frac{\mathbb{E}_{\lambda_0}[\pi_{\nu ij}]}{\pi_{\nu i}^\lambda \pi_{\nu j}^\lambda} - 1\right|\right], \tag{34}$$

which is nearly identical to Savitsky and Toth (2016) with the raw, $\pi_{\nu i}$, replaced by the modeled, $\pi_{\nu i}^\lambda = \mathbb{E}_{P_\theta}(\pi_{\nu i}^\kappa)$, where $\pi_{\nu i}^\kappa = \mathbb{E}_\kappa(\pi_{\nu i} | Y_{\nu i} = y_{\nu i})$ in the first term on the left of Equation (34). Under condition (A5), $\mathbb{E}_{\lambda_0}[\pi_{\nu ij}] \to \mathbb{E}_{\lambda_0}[\pi_{\nu i}\pi_{\nu j}] = \pi_{\nu i}^{\lambda_0}\pi_{\nu j}^{\lambda_0}$. Condition (A7) guarantees convergence to 1 of the ratio in the second term of Equation (34). ☐

Appendix D: Data preprocessing in Section 5

Here we describe how we obtained our dataset of $n = 6847$ participants that we used in analyses in Section 5. We downloaded the "Dietary Interview - Total Nutrient Intakes, First Day" and "Blood Pressure" datasets. These datasets, named DR1TOT_H.XPT and BPX_H.XPT, are available at https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Dietary&CycleBeginYear=2013 and `https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/BPX_H.htm`, respectively.

The first dataset includes the "Dietary day one sample weight" variable (called WTDRD1) and its documentation states that this variable should be

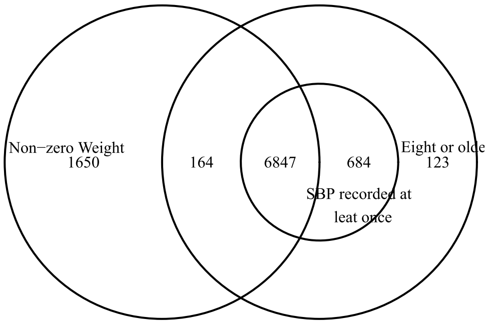

Fig 8. *Venn diagram the three sets of NHANES participants. The sets are with non-zero weights, 8 years or older and SBP recorded at least once.*

used as a sampling weight when analyzing it, as we did in our analyzes. The dataset contains the nutrient intake information 9813 NHANES participants. This dataset is derived from questionnaire data where each participant is asked questions on salt, amounts of food and beverages consumed during the 24-hour period prior to the interview (midnight to midnight). For more details see the dataset documentation. 1152 participants have sampling weights equal to zero leaving 9813-1152=8661 (the participants with weight 0 also had missing values in all their nutrient consumption data) participants in this analysis.

The second dataset includes three consecutive blood pressure readings. According to its documentation, when a blood pressure measurement is interrupted or incomplete, a fourth attempt may be made. In our analysis SBP was the average of the not missing values of these four reads. This dataset contains the information of the 9813 participants in the first dataset but only 7818 are eight years or older (by design their SBP was not measured for younger participants), from these 807 have sampling weights equal to zero leaving 7818-807=7011. Out of these 7011, 6847 at least one (out of four) SBP measure recorded. Figure 8 in an Appendix depicts the Venn diagram of the sets of NHANES participants with non-zero sampling weights, 8 years or older and SBP recorded at least once.

To obtain age and gender information of the participants we downloaded the dataset DEMO_H.XPT available at `https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2013`.

References

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37, 2016.

CDC-A. Centers for Disease Control (CDC). national health and nutrition examination survey. `https://www.cdc.gov/nchs/nhanes/index.htm`, 2016. Accessed: June 7, 2016.

CDC-B. Centers for Disease Control (CDC). modeling usual intake using dietary recall data, task 4: Key concepts about variance estimation methods in NHANES. `https://www.cdc.gov/nchs/tutorials/dietary/Advanced/ModelUsualIntake/Info4.htm`, 2016. Accessed: June 7, 2016.

CDC-D. Centers for Disease Control (CDC). key concepts about NHANES survey design: NHANES sampling procedure. `http://www.cdc.gov/nchs/tutorials/NHANES/SurveyDesign/SampleDesign/Info1.htm`, 2016. Accessed: June 7, 2016.

Qi Dong, Michael R. Elliott, and Trivellore E. Raghunathan. A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, 40(1):29–46, 2014.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000. MR1790007

David R Judkins. Fay's method for variance estimation. *Journal of Official Statistics*, 6(3):223, 1990.

D Krewski and JNK Rao. Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, pages 1010–1019, 1981. MR0628756

Tsuyoshi Kunihama, AH Herring, CT Halpern, and DB Dunson. Nonparametric bayes modeling with sample survey weights. *Statistics & Probability Letters*, 113:41–48, 2016. MR3480393

Roderick J. Little. To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556, 2004. MR2109316

Philip J McCarthy. Pseudo-replication: Half samples. *Revue de l'Institut International de Statistique*, pages 239–264, 1969.

Robert Nishihara, Iain Murray, and Ryan P. Adams. Parallel mcmc with generalized elliptical slice sampling. *J. Mach. Learn. Res.*, 15(1):2087–2112, January 2014. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2627435.2670318`. MR3231603

Danny Pfeffermann, Abba M Krieger, and Yosef Rinott. Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, pages 1087–1114, 1998. MR1666233

Danny Pfeffermann, Fernando Antonio Da Silva Moura, and Pedro Luis Do Nascimento Silva. Multi-level modelling under informative sampling. *Biometrika*, 93(4):943–959, 2006. MR2285081

J. N. K. Rao and Changbao Wu. Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society Series B*, 72(4):533–544, 2010. MR2758527

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Sur-*

vey Sampling (Springer Series in Statistics). Springer Science & Business Media, 2003. ISBN 0387406204. MR1140409

Terrance D Savitsky and Daniell Toth. Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10(1):1677–1708, 2016. MR3522657

Yajuan Si, Natesh S Pillai, Andrew Gelman, et al. Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10(3):605–625, 2015. MR3420817

Paul L Speckman and Dongchu Sun. Fully bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90(2):289–302, 2003. MR1986647

Mark Steffen, Carol Kuhle, Donald Hensrud, Patricia J Erwin, and Mohammad H Murad. The effect of coffee consumption on blood pressure and the development of hypertension: a systematic review and meta-analysis. *Journal of hypertension*, 30(12):2245–2254, 2012.