Bayesian Analysis (2020) **15**, Number 1, pp. 57–77

# Bayesian Estimation Under Informative Sampling with Unattenuated Dependence

Matthew R. Williams[*] and Terrance D. Savitsky[†]

**Abstract.** An informative sampling design leads to unit inclusion probabilities that are correlated with the response variable of interest. However, multistage sampling designs may also induce higher order dependencies, which are ignored in the literature when establishing consistency of estimators for survey data under a condition requiring asymptotic independence among the unit inclusion probabilities. This paper constructs new theoretical conditions that guarantee that the pseudo-posterior, which uses sampling weights based on first order inclusion probabilities to exponentiate the likelihood, is consistent not only for survey designs which have asymptotic factorization, but also for survey designs that induce residual or unattenuated dependence among sampled units. The use of the survey-weighted pseudo-posterior, together with our relaxed requirements for the survey design, establish a wide variety of analysis models that can be applied to a broad class of survey data sets. Using the complex sampling design of the National Survey on Drug Use and Health, we demonstrate our new theoretical result on multistage designs characterized by a cluster sampling step that expresses within-cluster dependence. We explore the impact of multistage designs and order based sampling.

**Keywords:** cluster sampling, stratification, survey sampling, sampling weights, Markov chain Monte Carlo.

**MSC 2010 subject classifications:** 62D05, 62G20.

## 1 Introduction

Bayesian formulations are increasingly popular for modeling hypothesized distributions with complicated dependence structures. The primary interest of the data analyst is to perform inference about a finite population generated from an unknown population generating distribution. Our set-up is where the observed data are collected as a sample taken from that finite population by a government statistical agency or private research organization under a complex sampling design distribution. The complex sampling design results in probabilities of inclusion (of units from the population into the observed sample) that are associated with some response variable of interest. This association could result in an observed data set consisting of units that are not independent and identically distributed. This association induces a correlation between the response variable of interest and the inclusion probabilities. Sampling designs that induce this correlation are termed, “informative”, and the balance of information in

[*] National Center for Science and Engineering Statistics, National Science Foundation, Alexandria, VA, USA, matthew.dunn.williams@gmail.com

[†] U.S. Bureau of Labor Statistics, Washington, DC, USA, Savitsky.Terrance@bls.gov

https://doi.org/10.1214/18-BA1143

the sample is different from that in the population. Failure to account for this dependence caused by the sampling design could bias estimation of parameters that index the joint distribution hypothesized to have generated the population (Holt et al., 1980). While emphasis is often placed on the first order inclusions probabilities (individual probabilities of selection), an "informative" design may also have other features such as clustering and stratification that use population information and impact higher order joint inclusion probabilities. The impact of these higher order terms are more subtle, but we will demonstrate that they can also impact bias and consistency.

In this paper we are presented with samples acquired under an unequally-weighted, informative sampling design and our goal is to perform inference on the population distribution (or model parameters) from the observed sample. Savitsky and Toth (2016) proposed an automated approach that formulates a sampling-weighted pseudo-posterior density by exponentiating each likelihood contribution by a sampling weight constructed to be inversely proportional to its marginal inclusion probability, $\pi_i$, for units, $i = 1, \ldots, n$, where $n$ denotes the number of units in the observed sample. They demonstrate that the pseudo-posterior produces asymptotically unbiased estimation of the population model estimated on the observed sample. Yet, they *restrict* the class of sampling designs to those where the pairwise dependencies among units attenuate to 0 in the limit of the population size, $N$, to guarantee frequentist consistency of the pseudo-posterior distribution estimated on the sample data, at the true population generating distribution. While some sampling designs will meet this restricted criterion, nearly *all* designs used, in practice, won't; for example, a two-stage clustered sampling design where the number of clusters increases proportional to the population size $N$, but the number of units within each cluster remain relatively fixed such that the dependence induced at the second stage of sampling never attenuates to 0. Another common set of example survey sampling designs outside of the restricted class defined by the current literature are those which sample households as clusters in one stage and, next, sample individuals within those households in a following stage. Despite the current lack of theoretical results demonstrating consistency of the pseudo-posterior for these multistage sampling designs, the pseudo-posterior performs well (provides unbiased estimation of population model parameters), in practice.

This work provides new, relaxed theoretical conditions that guarantee consistency of the pseudo-posterior estimator for a much broader class of multistage sampling designs characterized by cluster steps. A cluster step groups units in a target population; for example, a geographic grouping is often done for convenience and cost to administer the survey by first sampling a geographic region followed by the sampling of units within selected regions. A multistage sampling design induces within-cluster dependence among the sampled units taken from the population and that dependence does not attenuate as the sample or population size grows. Yet, we prove that consistency is guaranteed for these dependence-inducing survey sampling designs. The practical significance of our innovation is that our expanded class of sampling designs under which our theoretical result guarantees frequentist consistency now includes nearly all commonly-used survey sampling designs by government statistical agencies and research organizations.

A second contribution of this paper is the design and implementation of a simulation study that very clearly illustrates our theoretical results. We, first, construct a sampling design from outside the new, broader class where there is no restriction on dependence among units and show that the sampling-weighted pseudo-posterior does not contract on the true generating distribution. We, next, make a minor change to this design by embedding the dependent units within clusters (or strata) where the dependence is unrestricted *within* each stratum, but attenuates to 0 *between* strata. This second design is now a member of the broader class described by our revised theory under which consistency is guaranteed and our simulation result shows that consistency is, indeed, achieved.

## 1.1 Examples

We next outline examples of commonly-used multistage sampling designs, including that for the National Survey on Drug Use and Health that motivate our simulation study and application modeling.

### Example 1: The Current Expenditure Survey

The Current Expenditure (CE) survey is administered to U.S. households by the U.S. Bureau of Labor Statistics for the purpose of determining the amount of spending for a broad collection of goods and service categories and it serves as the main source used to construct the basket of goods later used to formulate the Consumer Price Index. The CE employs a multistage sampling design that draws clusters of core-based statistical areas (CBSAs), such as metropolitan and micropolitan areas, from which census blocks and, ultimately, households are sampled. Economists desire to model the propensity to purchase a variety of goods and services. There is a dependence among households within census block that doesn't attenuate as the sample size increases. The revised theoretical conditions of this paper demonstrate that asymptotically unbiased estimation is achieved for the pseudo-posterior formulation under such designs - where dependence is concentrated within clusters.

### Example 2: The National Survey on Drug Use and Health

Our simulation study and application in the sequel are both motivated by the National Survey on Drug Use and Health (NSDUH), sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA). NSDUH is the primary source for statistical information on illicit drug use, alcohol use, substance use disorders (SUDs), mental health issues, and their co-occurrence for the civilian, non institutionalized population of the United States. The NSDUH employs a multistage state-based design (Morton et al., 2016), with the earlier stages defined by geography within each state in order to select households (and group quarters) nested within these geographically-defined primary sampling units (PSUs). The sampling frame was stratified implicitly by sorting the first-stage sampling units by a CBSA and socioeconomic status indicator and by the percentage of the population that is non-Hispanic and white. First stage units (census tracts) were then selected with probability proportionate to a composite size measure based on age groups. This selection was performed 'systematically' along the sort order gradient. Second and third stage units (census block groups and census blocks) were sorted geographically and selected with probability proportionate to size (PPS) sequentially along the sort order. Fourth stage dwelling units (DU) were selected systematically with equal probability, selecting every $k^{th}$ DU after a random starting point. Within households, 0, 1 or 2 individuals were selected with unequal probabilities depending on age with youth (age 12-17) and young adults (age 18-25) over-sampled.

This paper provides conditions for asymptotic consistency for designs like the NSDUH, which are characterized by:

- Cluster sampling, such as selecting only one unit per cluster, or selecting multiple individuals from a dwelling unit.
- Population information used to sort sampling units along gradients.

Both features are common, in practice, and create pairwise sampling dependencies that do not attenuate even if the population grows. The consistency of estimators under these sampling designs are not addressed in the literature. For example, we will examine the relationship between depression and smoking. Cigarette use and depression vary by age, metropolitan vs. non-metropolitan status, education level, and other demographics (Center for Behavioral Health Statistics and Quality, 2015b,a). Both smoking and depression have the potential to cluster geographically and within dwelling units, since these related demographics may cluster. Yet the current literature, such as in Savitsky and Toth (2016), is silent on the issue of non-ignorable clustering that may be informative (i.e. related to the response of interest). The results presented in this work establish conditions for a wide variety of survey designs and provide a theoretical justification that this relationship can be estimated consistently even under a complex multistage design such as the NSDUH.

## 1.2 Review of Methods to Account for Dependent Sampling

For consistency results, assumptions of approximate or asymptotic independence of sample selection (or factorization of joint inclusion probabilities into a product of individual inclusion probabilities) are ubiquitous. For example, Isaki and Fuller (1982) assume asymptotic factorization to demonstrate the consistency of the Horvitz-Thompson estimator and related regression estimators. More recently, Toth and Eltinge (2011) used a similar assumption to demonstrate consistency of survey-weighted regression trees and Savitsky and Toth (2016) used it to show consistency of a survey-weighted pseudo-posterior.

Chambers and Skinner (2003, Ch.2) review the construction of a sample likelihood using a Bayes rule expression for the population $U$ likelihood defined on the units in the sample $s$, $f_s(y) = f_U(y|I = 1)$ (similar to Pfeffermann et al. 1998). They explicitly state the assumption that the “sample inclusion for any particular population unit is independent of that for any other unit (and is determined by the outcome of a zero-one random variable, $I$)”. The further assumption of independence of the population units stated in

Pfeffermann et al. (1998) means that weighting each likelihood contribution multiplied together in the sample is an approximation of the likelihood for the $N$ population units.

Pfeffermann et al. (1998) maintains the assumption of unconditional independence of the population units, but defines two classes of sampling designs: (1) The first class is independent, with replacement sampling, so the sample inclusions are all independent. (2) The second class is some selected with replacement designs that are asymptotically independent. Chambers and Skinner (2003) discuss the pseudo-likelihood (and cite Kish and Frankel, 1974, Binder, 1983 and Godambe and Thompson, 1986) for estimation via a weighted score function. They assume that the correlation between inclusion indicators has an *expected value* of 0, where the expectation is with respect to the population generating distribution. We note that they do not assume this correlation to be exactly equal to 0. However this condition still appears to be more restrictive than that of asymptotic factorization in which deviations from factorization shrink to 0 at a rate inverse to the population size $N$: $\mathcal{O}(N^{-1})$.

The assumptions above are relied on to show consistency. However in practice, approximate sampling independence is only assumed for the first stage or primary sampling units (PSUs), with dependence between secondary units within these clusters commonly assumed. This setup is the defacto approach for design-based variance estimation (for example, see Heeringa et al., 2010, Ch.3 and Rao et al., 1992) and is used in all the major software packages for analyzing survey data. One goal of the current work is to reconcile this discrepancy by extending the class of designs for which consistency results are available to cover designs seen in practice such as those for which design-based variance estimation strategies already exist.

We focus on extending the results of the survey-weighted pseudo-posterior method of Savitsky and Toth (2016) which provides for flexible modeling of a very wide class of population generating models. By refining and relaxing the conditions on factorization, we expand results to include many common sampling designs. These conditions for the sampling designs can be applied to generalize many of the other consistency results mentioned above. There are some population models of interest for which marginal inclusion probabilities may not be sufficient and pairwise inclusion probabilities and composite likelihoods can be used to achieve consistent results (Yi et al., 2016; Williams and Savitsky, 2018). However, Williams and Savitsky (2018) demonstrate that both a very specific population model (for example conditional behavior of spouse-spouse pairs within households) and specific sample design (differential selection of pairs of individuals within a household related to outcome) are needed for marginal weights to lead to bias. In the usual setting of inference on a population of individuals (rather than on a population of joint relationships *within* households), pairwise weights and marginal weights are numerically similar, converging to one another for moderate sample sizes. The theory presented in the current work also clarifies why both approaches lead to consistent results. Furthermore, the current work also applies when individual units are mutually exclusive; for example, only selecting one individual from a household to the exclusion of all others. Such designs are not covered by the composite likelihood with pairwise weights approach, which require non-zero joint inclusion probabilities.

The remainder of this work proceeds as follows: In Section 2 we briefly review the pseudo-posterior approach to account for informative sampling via the exponentiation of the likelihood with sampling weights. Our main result, presented in Section 3, provides the formal conditions controlling for sampling dependence. In Section 4, we provide two simulations. We first demonstrate consistency for a multistage survey design analogous to the NSDUH. We next create a pathological design based on sorting. The design violates our assumptions for sampling dependence and estimates fail to converge. However, we show that this design will lead to consistency if embedded within stratified or clustered designs. Lastly, we revisit the NSDUH with two simple examples (Section 5) and provide some conclusions (Section 6).

## 2 Pseudo-Posterior Estimator to Account for Informative Sampling

We briefly review the pseudo-likelihood and associated pseudo-posterior as constructed in Savitsky and Toth (2016) and revisited by Williams and Savitsky (2018).

Suppose there exists a Lebesgue measurable population-generating density, $\pi(y|\boldsymbol{\lambda})$, indexed by parameters, $\boldsymbol{\lambda} \in \Lambda$. Let $\delta_i \in \{0,1\}$ denote the sample inclusion indicator for units $i = 1, \ldots, N$ from the population. The density for the observed sample is denoted by, $\pi(y_o|\boldsymbol{\lambda}) = \pi(y|\delta = 1, \boldsymbol{\lambda})$, where "$o$" indicates "observed".

The plug-in estimator for posterior density under the analyst-specified model for $\boldsymbol{\lambda} \in \Lambda$ is

$$\hat{\pi}\left(\boldsymbol{\lambda}|\mathbf{y}_o, \tilde{\mathbf{w}}\right) \propto \left[\prod_{i=1}^{n} p\left(y_{o,i}|\boldsymbol{\lambda}\right)^{\tilde{w}_i}\right] \pi\left(\boldsymbol{\lambda}\right), \tag{1}$$

where $\prod_{i=1}^{n} p(y_{o,i}|\boldsymbol{\lambda})^{\tilde{w}_i}$ denotes the pseudo-likelihood for observed sample responses, $\mathbf{y}_o$. The joint prior density on model space assigned by the analyst is denoted by $\pi(\boldsymbol{\lambda})$. The sampling weights, $\{\tilde{w}_i \propto 1/\pi_i\}$, are inversely proportional to unit inclusion probabilities and normalized to sum to the sample size, $n$. Let $\hat{\pi}$ denote the noisy approximation to posterior distribution, $\pi$, based on the data, $\mathbf{y}_o$, and sampling weights, $\{\tilde{\mathbf{w}}\}$, confined to those units *included* in the sample, $S$.

## 3 Consistency of the Pseudo-Posterior Estimator

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 $\nu$ increases. Suppose that $\mathbf{X}_{\nu,1}, \ldots, \mathbf{X}_{\nu,N_\nu}$ are independently generated from an unknown distribution $P_0$, (with density, $p_0$) defined on the sample space, $(\mathcal{X}, \mathcal{A})$. A sampling design distribution, $P_\nu$, is defined by placing a *known* distribution on a vector of inclusion indicators, $\boldsymbol{\delta}_\nu = (\delta_{\nu 1} \in \{0,1\}, \ldots, \delta_{\nu N_\nu} \in \{0,1\})$, linked to the units comprising the population, $U_\nu$. The (survey) sampling distribution is subsequently used to take an *observed* random sample of size $n_\nu \leq N_\nu$. Our conditions needed for the main result employ known marginal unit inclusion probabilities, $\pi_{\nu i} = \Pr\{\delta_{\nu i} = 1\}$ for all $i \in U_\nu$ and the second order pairwise probabilities, $\pi_{\nu ij} = \Pr\{\delta_{\nu i} = 1 \cap \delta_{\nu j} = 1\}$ for $i, j \in U_\nu$, which are obtained from the joint distribution over $(\delta_{\nu 1},\ldots,\delta_{\nu N_\nu})$. We denote the (survey) sampling distribution by $P_\nu$. A common sampling design constructs the $\pi_{\nu i}$ to be proportional to the $\mathbf{X}_{\nu i}$, such that units, $i$, with larger values for $\mathbf{X}_{\nu i}$ are more likely to be included in the sample. This type of sampling design is referred to as probability proportional to size (PPS). For example, the U.S. Bureau of Labor statistics administers the Current Employment Statistics (CES) survey of business establishments in a geography and industry to assess the total employment for that geography and industry. Larger establishments with more employees are assigned a higher probability of inclusion into the survey because these larger establishments drive more of the variance in the estimator, so that including more larger establishments will produce a more efficient, lower variance estimator of total employment.

The asymptotics under our construction is controlled by $\nu$. We fix a $\nu$, construct an associated finite population of size, $N_\nu$, generate random variables $\mathbf{X}_{\nu 1},\ldots,\mathbf{X}_{\nu N_\nu}\sim P_0$, construct unit marginal sample inclusion probabilities, $(\pi_{\nu 1},\ldots,\pi_{\nu N})$ under $P_\nu$ and then draw a sample, $\{1,\ldots,n_\nu\}$ from that population.

Under informative sampling, the inclusion probabilities are formulated to depend on the finite population data values, $\mathbf{X}_{N_\nu}=(\mathbf{X}_1,\ldots,\mathbf{X}_{N_\nu})$. Information from the population is used to determine size measures for unequal selection $\pi_{\nu i}$ and used to establish clustering and stratification which determine joint inclusions probabilities $\pi_{\nu ij}$. Since the balance of information is different between the population and a resulting sample, a posterior distribution for $(\mathbf{X}_1\delta_{\nu 1},\ldots,\mathbf{X}_{N_\nu}\delta_{\nu N_\nu})$ that ignores the distribution for $\boldsymbol{\delta}_\nu$ will not lead to consistent estimation.

Our task is to perform inference about the population generating distribution, $P_0$, using the observed data taken under an informative sampling design. We account for informative sampling by "undoing" the sampling design with the weighted estimator,

$$p^\pi\left(\mathbf{X}_i\delta_{\nu i}\right):=p\left(\mathbf{X}_i\right)^{\delta_{\nu i}/\pi_{\nu i}},\ i\in U_\nu, \tag{2}$$

which weights each density contribution, $p(\mathbf{X}_i)$, by the inverse of its marginal inclusion probability. This approximation for the population likelihood produces the associated pseudo-posterior,

$$\Pi^\pi\left(B|\mathbf{X}_1\delta_{\nu 1},\ldots,\mathbf{X}_{N_\nu}\delta_{\nu N_\nu}\right)=\frac{\int_{P\in B}\prod_{i=1}^{N_\nu}\frac{p^\pi}{p_0^\pi}(\mathbf{X}_i\delta_{\nu i})d\Pi(P)}{\int_{P\in\mathcal{P}}\prod_{i=1}^{N_\nu}\frac{p^\pi}{p_0^\pi}(\mathbf{X}_i\delta_{\nu i})d\Pi(P)}, \tag{3}$$

that we use to achieve our required conditions for the rate of contraction of the pseudo-posterior distribution on $P_0$. We note that both $P$ and $\boldsymbol{\delta}_\nu$ are random variables defined on the space of measures ($\mathcal{P}$ and $B\subseteq\mathcal{P}$) and the distribution, $P_\nu$, governing all possible samples, respectively. An important condition on $P_\nu$ formulated in Savitsky and Toth (2016) that guarantees contraction of the pseudo-posterior on $P_0$ restricts pairwise inclusion dependencies to asymptotically attenuate to 0. This restriction narrows the class of sampling designs for which consistency of a pseudo-posterior based on marginal inclusion probabilities may be achieved. We will replace their condition that requires marginal factorization of all pairwise inclusion probabilities with a less restrictive condition allowing for non-factorization for a small partition of pairwise inclusion probabilities. This expands the allowable class of sampling designs under which frequentist consistency may be guaranteed. We assume measurability for the sets on which we compute prior, posterior and pseudo-posterior probabilities on the joint product space, $\mathcal{X} \times \mathcal{P}$. For brevity, we use the superscript, $\pi$, to denote the dependence on the known sampling probabilities, $\{\pi_{\nu ij}\}_{i,j \in U_\nu}$; for example,

$$\Pi^{\pi}\left(B|\mathbf{X}_1\delta_{\nu 1},\ldots,\mathbf{X}_{N_\nu}\delta_{\nu N_\nu}\right) := \Pi\left(B|\left(\mathbf{X}_1\delta_{\nu 1},\ldots,\mathbf{X}_{N_\nu}\delta_{\nu N_\nu}\right),\{\pi_{\nu ij} : i,j \in U_\nu\}\right).$$

Our main result is achieved in the limit as $\nu \uparrow \infty$, under the countable set of successively larger-sized populations, $\{U_\nu\}_{\nu \in \mathbb{Z}^+}$. We define the associated rate of convergence notation, $a_\nu = \mathcal{O}(b_\nu)$, to denote $|a_\nu| \leq M|b_\nu|$ for a constant $M > 0$.

## 3.1 Empirical Process Functionals

We employ the empirical distribution approximation for the joint distribution over population generation and the draw of an informative sample that produces our observed data to formulate our results. Our empirical distribution construction follows Breslow and Wellner (2007) and incorporates inverse inclusion probability weights, $\{1/\pi_{\nu i}\}_{i=1,\ldots,N_\nu}$, to account for the informative sampling design,

$$\mathbb{P}^{\pi}_{N_\nu} = \frac{1}{N_v}\sum_{i=1}^{N_\nu}\frac{\delta_{\nu i}}{\pi_{\nu i}}\delta\left(\mathbf{X}_i\right), \tag{4}$$

where $\delta(\mathbf{X}_i)$ denotes the Dirac delta function, with probability mass 1 on $\mathbf{X}_i$ and we recall that $N_\nu = |U_\nu|$ denotes the size of the finite population. This construction contrasts with the usual empirical distribution, $\mathbb{P}_{N_\nu} = \frac{1}{N_v}\sum_{i=1}^{N_\nu}\delta(\mathbf{X}_i)$, used to approximate $P \in \mathcal{P}$, the distribution hypothesized to generate the finite population, $U_\nu$.

We follow the notational convention of Ghosal et al. (2000) and define the associated expectation functionals with respect to these empirical distributions by $\mathbb{P}^{\pi}_{N_\nu} f = \frac{1}{N_\nu}\sum_{i=1}^{N_\nu}\frac{\delta_{\nu i}}{\pi_{\nu i}} f(\mathbf{X}_i)$. Similarly, $\mathbb{P}_{N_\nu} f = \frac{1}{N_\nu}\sum_{i=1}^{N_\nu} f(\mathbf{X}_i)$. Lastly, we use the associated centered empirical processes, $\mathbb{G}^{\pi}_{N_\nu} = \sqrt{N_\nu}(\mathbb{P}^{\pi}_{N_\nu} - P_0)$ and $\mathbb{G}_{N_\nu} = \sqrt{N_\nu}(\mathbb{P}_{N_\nu} - P_0)$.

The sampling-weighted, (average) pseudo-Hellinger distance between distributions, $P_1, P_2 \in \mathcal{P}$, $d^{\pi,2}_{N_\nu}(p_1,p_2) = \frac{1}{N_\nu}\sum_{i=1}^{N_\nu}\frac{\delta_{\nu i}}{\pi_{\nu i}} d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i))$, where $d(p_1,p_2) = [\int(\sqrt{p_1}-\sqrt{p_2})^2 d\mu]^{\frac{1}{2}}$ (for dominating measure, $\mu$). We need this empirical average distance metric because the observed (sample) data drawn from the finite population under $P_\nu$ are no longer independent. The associated non-sampling Hellinger distance is specified with, $d^2_{N_\nu}(p_1,p_2) = \frac{1}{N_\nu}\sum_{i=1}^{N_\nu} d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i))$.

## 3.2 Main Result

We proceed to construct associated conditions and a theorem that contain our main result on the consistency of the pseudo-posterior distribution under a broader class of informative sampling designs at the true generating distribution, $P_0$. This approach follows the main in-probability convergence result of Savitsky and Toth (2016) which extends Ghosal and van der Vaart (2007) by adding new conditions that restrict the distribution of the informative sampling design. Instead of the standard asymptotic factorization condition, we provide two alternative conditions which allow for residual dependence between sampling units:

Suppose we have a sequence, $\xi_{N_\nu} \downarrow 0$ and $N_\nu \xi_{N_\nu}^2 \uparrow \infty$ and $n_\nu \xi_{N_\nu}^2 \uparrow \infty$ as $\nu \in \mathbb{Z}^+ \uparrow \infty$ and any constant, $C > 0$,

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

$$\sup_{\xi > \xi_{N_\nu}} \log N\left(\xi/36, \{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}\left(P, P_0\right) < \xi\}, d_{N_\nu}\right) \leq N_\nu \xi_{N_\nu}^2$$

**(A2)** (Size of space)

$$\Pi\left(\mathcal{P} \backslash \mathcal{P}_{N_\nu}\right) \leq \exp\left(-N_\nu \xi_{N_\nu}^2 \left(2(1+2C)\right)\right)$$

**(A3)** (Prior mass covering the truth)

$$\Pi\left(P : -P_0 \log \frac{p}{p_0} \leq \xi_{N_\nu}^2 \cap P_0 \left[\log \frac{p}{p_0}\right]^2 \leq \xi_{N_\nu}^2\right) \geq \exp\left(-N_\nu \xi_{N_\nu}^2 C\right)$$

**(A4)** (Non-zero Inclusion Probabilities)

$$\sup_\nu \left[\frac{1}{\min\limits_{i \in U_\nu} |\pi_{\nu i}|}\right] \leq \gamma, \text{ with } P_0\text{−probability } 1.$$

**(A5.1)** (Growth of dependence is restricted)
For every $U_\nu$ there exists a binary partition $\{S_{\nu 1}, S_{\nu 2}\}$ of the set of all pairs $S_\nu = \{\{i, j\} : i \neq j \in U_\nu\}$ such that

$$\limsup_{\nu \uparrow \infty} |S_{\nu 1}| = \mathcal{O}(N_\nu),$$

and

$$\limsup_{\nu \uparrow \infty} \max_{i,j \in S_{\nu 2}} \left|\frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}} - 1\right| = \mathcal{O}(N_\nu^{-1}), \text{ with } P_0\text{−probability } 1$$

such that for some constants, $C_4, C_5 > 0$ and for $N_\nu$ sufficiently large,

$$|S_{\nu 1}| \leq C_4 N_\nu,$$

and

$$N_\nu \sup_\nu \max_{i,j \in S_{\nu 2}} \left|\frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}} - 1\right| \leq C_5.$$

**(A5.2)** (Dependence restricted to countable blocks of bounded size)
For every $U_\nu$ there exists a partition $\{B_1, \ldots, B_{D_\nu}\}$ of $U_\nu$ with $D_\nu \leq N_\nu$, $\lim_{\nu\uparrow\infty} D_\nu = \mathcal{O}(N_\nu)$, and the maximum size of each subset is bounded:

$$1 \leq \sup_\nu \max_{d \in 1,\ldots,D_\nu} |B_d| \leq C_4.$$

Such that the set of all pairs $S_\nu = \{\{i,j\} : i \neq j \in U_\nu\}$ can be partitioned into $S_{\nu 1} = \{\{i,j\} : i \neq j \in B_d, d \in \{1,\ldots,D_\nu\}\}$ and
$S_{\nu 2} = \{\{i,j\} : i \in B_d \cap j \notin B_d, d \in \{1,\ldots,D_\nu\}\}$ with

$$\limsup_{\nu\uparrow\infty} \max_{i,j \in S_{\nu_2}} \left| \frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}} - 1 \right| = \mathcal{O}(N_\nu^{-1}), \text{ with } P_0\text{−probability } 1$$

such that for some constant, $C_5 > 0$,

$$N_\nu \sup_\nu \max_{i,j \in S_{\nu_2}} \left| \frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}} - 1 \right| \leq C_5, \text{ for } N_\nu \text{ sufficiently large.}$$

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

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

The first three conditions are the same as Ghosal and van der Vaart (2007). They restrict the growth rate of the model space (e.g., of parameters) and require prior mass to be placed on an interval containing the true value. 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. Condition (A6) ensures that the observed sample size, $n_\nu$, limits to $\infty$ along with the size of the partially-observed finite population, $N_\nu$, such that the variation of information about the population expressed in realized samples is controlled.

Savitsky and Toth (2016) rely on asymptotic factorization for all pairwise inclusion probabilities. Their (A.5) condition is a conservative approach to establish a finite upper bound for the un-normalized posterior mass assigned to those models, $P$, at some minimum distance from the truth, $P_0$. They require all terms, a set of size $\mathcal{O}(N_\nu^2)$, to factorize with the maximum deviation term shrinking at a rate of $\mathcal{O}(N_\nu^{-1})$, since there are $N^2$ terms divided by $N$ (inherited from an empirical process).

Although their condition guarantees the $L_1$ contraction result, it defines an overly narrow class of sampling designs under which this guaranteed result holds. As discussed in the introduction, multistage household survey designs are not members of this allowed class because the within household dependency does not attenuate for a set of pairs of size $\mathcal{O}(N_\nu)$. We replace their (A5) with (A5.1), which allows up to $\mathcal{O}(N_\nu)$ pairwise terms to not factor, such that there remains a residual dependence. We show in the Supplementary Material (Williams and Savitsky, 2019) that the contraction result may, nevertheless, be guaranteed under this condition as each of the non-factoring terms has an $\mathcal{O}(1)$ bound. The implication of our condition is that we have constructed a wider class of sampling designs that includes those from Savitsky and Toth (2016), in addition to the multistage cluster designs for fixed cluster sizes.

Our condition (A5.2) is a special case of (A5.1) specified for cluster designs where the number of units per cluster is bounded by a constant, which encompasses the multistage NSDUH household design from which we draw our application data set. We walk from (A5.1) to (A5.2) by constructing $S_{\nu 1}$ through a collection of clusters $(B_{\nu 1}, \ldots, B_{\nu D_\nu})$, where the size $|B_{\nu d}|$ is bounded from above. Sampling dependence within each cluster $B_{\nu d}$ is unrestricted, while dependence across clusters must asymptotically factor.

**Theorem 1.** *Suppose conditions (A1)–(A6) hold. Then for sets $\mathcal{P}_{N_\nu} \subset \mathcal{P}$, constants, $K > 0$, and $M$ sufficiently large,*

$$\mathbb{E}_{P_0,P_\nu} \Pi^\pi \left(P : d^\pi_{N_\nu}\left(P, P_0\right) \geq M\xi_{N_\nu} | \mathbf{X}_1\delta_{\nu 1}, \ldots, \mathbf{X}_{N_\nu}\delta_{\nu N_\nu}\right) \leq \frac{16\gamma^2\left[\gamma\mathbf{C_2} + C_3\right]}{\left(Kf + 1 - 2\gamma\right)^2 N_\nu \xi^2_{N_\nu}} + 5\gamma \exp\left(-\frac{K n_\nu \xi^2_{N_\nu}}{2\gamma}\right), \tag{5}$$

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

*Proof.* The expectation is taken with respect to the joint distribution, $(P_0, P_\nu)$, over population generation and the taking of a survey sample from that population. The proof follows exactly that in Savitsky and Toth (2016) where we bound the numerator (from above) and the denominator (from below) of the expectation with respect to the joint distribution of population generation and the taking of a sample of the pseudo-posterior mass placed on the set of models, $P$, at some minimum pseudo-Hellinger distance from $P_0$. We reformulate one of the enabling lemmas of Savitsky and Toth (2016), which we present in the Supplementary Material, where the reliance on (their) condition (A5) requiring asymptotic factoring of pairwise unit inclusion probabilities is here replaced by condition (A5.1) that allows for non-factorization of a subset of pairwise inclusion probabilities. □

As noted in Savitsky and Toth (2016), the rate of convergence is decreased for a sampling distribution, $P_\nu$, that expresses a large variance in unit pairwise inclusion probabilities such that $\gamma$ will be relatively larger. Samples drawn under a design that expresses a large variability in the first order sampling weights will express more dispersion in their information relative to a simple random sample of the underlying finite population. We construct $C_3 = C_5 + 1$ and $C_2 = C_4 + 1$. Under the more restrictive condition (A5) of Savitsky and Toth (2016), our constant $C_4 = 0$ and thus $C_2 = 1$.

# 4 Simulation Examples

We construct a population model to address our inferential interest of a binary outcome $y$. Many health related outcomes such as substance use are binary. For simplicity of demonstration we consider a linear predictor $\mu = \boldsymbol{X\beta}$. However more complex models can be used.

$$y_i \mid \mu_i \overset{\text{ind}}{\sim} Bern\left(\theta_i = F_l(\mu_i)\right),\ i = 1, \ldots, N \tag{6}$$

where $\theta_i = P(y_i = 1)$ and $F_l$ is the cumulative distribution function for the logistic distribution. We let $\mu$ depend on two predictors $x_1$ and $x_2$. The variable $x_1$ represents the observed information available for analysis, whereas $x_2$ represents information available for sampling, which is either ignored or not available for analysis. The $x_1$ and $x_2$ distributions are $\mathcal{N}(0,1)$ and $\mathcal{E}(r = 1/5)$ with rate $r$, where $\mathcal{N}(\cdot)$ and $\mathcal{E}(\cdot)$ represent normal and exponential distributions, respectively. The size measure used for sample selection is $\tilde{\boldsymbol{x}}_2 = \boldsymbol{x}_2 - \min(\boldsymbol{x}_2) + 1$.

$$\boldsymbol{\mu} = -1.88 + 1.0\boldsymbol{x}_1 + 0.5\boldsymbol{x}_2$$

where the intercept was chosen such that the median of $\mu$ is approximately 0, therefore the median of $\boldsymbol{\theta} = F_l(\boldsymbol{\mu})$ is approximately 0.5 (Results for small $\theta_i$ are similar but take larger sample sizes $n$ to converge as $\theta \rightarrow 0$).

Even though the population response $y$ was simulated with $\mu = f(x_1, x_2)$, we estimate the marginal model at the population level for $\mu = f(x_1)$. This exclusion of $x_2$ is analogous to the situation in which an analyst does not have access to all the sample design information and ensures that our sampling design instantiates informativeness (where $y$ is correlated with the selection variable, $x_2$, that defines inclusion probabilities). In particular, we estimate the models under each of several sample design scenarios and compare the population fitted models, $\mu = f(x_1)$, to those from the samples.

We formulate the logarithm of the sampling-weighted pseudo-likelihood for estimating $(\boldsymbol{\mu}, \lambda)$ from our observed data for the $n \leq N$ sampled units,

$$\begin{aligned}\log\left[\prod_{i=1}^{n} p\left(y_i \mid x_{1i}, \beta_0, \beta_1\right)^{\tilde{w}_i}\right] &= \sum_{i=1}^{n} \tilde{w}_i \log p\left(y_i \mid x_{1i}, \beta_0, \beta_1\right) \\ &= \sum_{i=1}^{n} \tilde{w}_i y_i \log(F_l(\beta_0 + x_{1i}\beta_1)) \\ &\quad + \tilde{w}_i(1 - y_i)\log(1 - F_l(\beta_0 + x_{1i}\beta_1)),\end{aligned} \tag{7}$$

where $\theta_i = F_l(\mu_i)$, $\mu_i = \beta_0 + x_{1i}\beta_1$, and the sampling weights, $\tilde{w}_i$ are normalized such that the sum of the weights equals the sample size $\sum_{i=1}^{n} \tilde{w}_i = n$.

Finally, we estimate the joint posterior distribution using (7), using the NUTS Hamiltonian Monte Carlo algorithm implemented in Stan (Carpenter, 2015). Example code is provided in the Supplementary Material. Prior distributions for $\beta$ were chosen as $\propto 1$. A sensitivity analysis using heavy-tailed proper priors in the $t$ family yielded essentially the same results (not shown).

## 4.1 Multistage Cluster Designs

We begin by abstracting the five-stage, geographically-indexed NSDUH sampling design (Morton et al., 2016) to a simpler, three stage design of {area segment, household, individual} that we use to draw samples from a synthetic population in a manner that still generalizes to the NSDUH (and similar multistage sampling designs where the number of last stage units does not grow with overall population size). We simulate a population of N = 6000, with 200 primary sampling units (PSUs) each containing 10 households (HHs) which each contain 3 individuals with independent responses $y_i$.

For the simulation, the number of selected PSUs was varied $K \in \{10, 20, 40, 80, 160\}$, the number of selected HHs within each PSU was fixed at 5, and the number of selected individuals within each HH was 1. Each setting was repeated $M = 200$ times. Details for the selection at each stage follows:

1. For each PSU indexed by $k$, an aggregate size measure $X_{2,k} = \sum_{ij} x_{2,ij|k}$ was created summing over all individuals $i$ and HHs $j$ in PSU $k$. PSUs are then selected proportional to this size measure based on Brewer's PPS algorithm (Brewer, 1975).

2. Once PSUs are selected, for each HH within the selected PSUs indexed by $j$, an aggregate size measure $X_{2,j|k} = \sum_i x_{2,i|jk}$ was created summing over all individuals $i$ within each HH in the selected PSUs. HHs are selected independently across PSUs. Within each PSU, HHs are selected systematically with equal probability by first sorting on $X_{2,j|k}$ and then selecting a random starting point.

3. Within each selected HH, a single person is selected with probability proportional to size $x_{2,i|jk}$.

The nested structure of the sampling induces asymptotic independence between PSUs. Within PSUs, the systematic sampling of HHs creates a block of non-attenuating dependence between households. Likewise, the sampling of only one person within each HH creates a joint dependence $\pi_{ii'|jk} = 0$ between individuals within the same HH. Therefore, non-factorization of the second order inclusions remains within each PSU (see Figure 1). Figure 2 compares the bias and mean square error (MSE) for estimation with equal weights (blue) and inverse probability weights (red). As expected, the sampling weights remove bias and lead to convergence, since the non-factoring pairwise inclusion probabilities are of $\mathcal{O}(N)$.

## 4.2 Dependent Sampling of First Stage Units

We now use the same population response model and distributions for $y$, $x_1$, and $x_2$ but consider the case of single stage sampling designs where the sample size is half the population (i.e. a partition of size $N/2$). In particular, we construct a design with second order dependence that grows $\mathcal{O}(N^2)$ and demonstrate that estimates for this design fail to converge. However, with slight modifications, the design can be altered into $\mathcal{O}(N)$ dependence and does demonstrate convergence, as predicted by the theory.

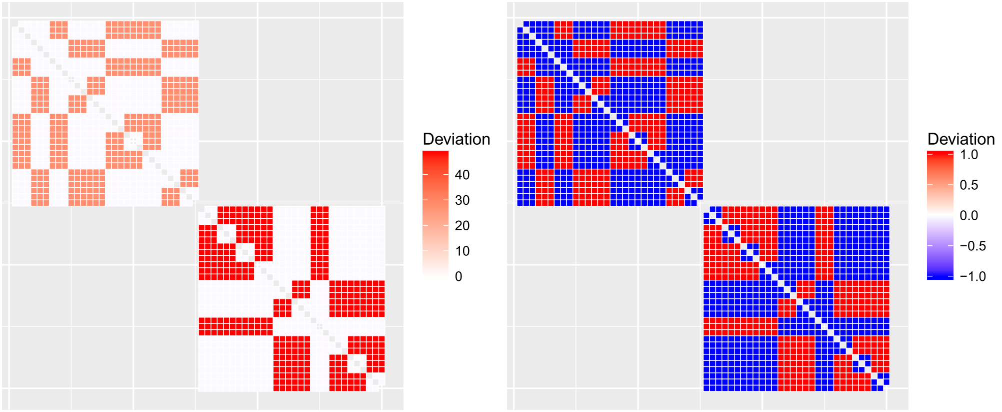

Figure 1: Matrix $\{i, j\}$ of deviations from factorization $(\pi_{ij}/(\pi_i\pi_j) - 1)$ for two PSUs (out of a population of 200) from the three stage sample design. Each PSU contains 10 HHs, which each contain 3 persons. Magnitude (left) and sign (right) of deviations. Empty cells correspond to 0 deviation (factorization). Created with 'ggplot2' (Wickham, 2009).

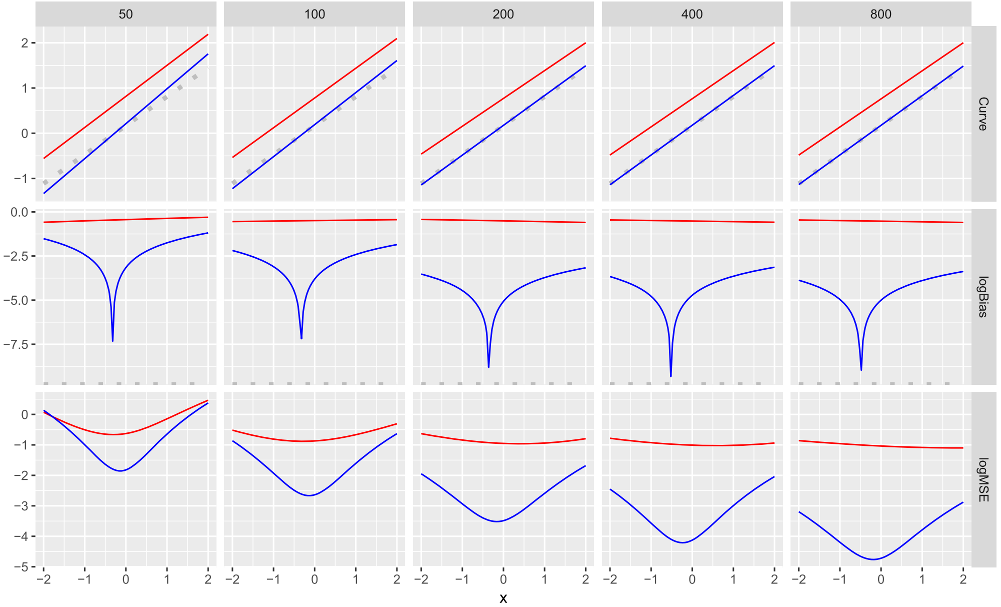

Figure 2: The marginal estimate of $\mu = f(x_1)$ under a linear model and three-stage sampling design. Compares the whole population curve (broken grey) to the sample with equal weights (red/light grey), and inverse probability weights (blue/dark grey). Top to bottom: estimated curve, log of absolute bias, log of mean square error. Left to right: doubling of sample size (50 to 800). Created with 'ggplot2' (Wickham, 2009).

One simple way to create an informative design is to use the size measure $\tilde{\boldsymbol{x}}_2$ to sort the population. Partition the population $U$ into a "high" ($U_1$) group with the top $N/2$ and a "low" ($U_2$) group with the bottom $N/2$. This partition rule leads to an outcome space with only two possible samples of size $N/2$: $U_1$ and $U_2$. For simplicity, assume an equal probability of selection of $1/2$. Then it follows that $\pi_i = 1/2$, for all $i \in 1, \ldots, N$, and $\pi_{ij} = 1/2$ if $i, j \in U_k$, for $k = 1, 2$ but $0$ if $i \in U_k$ and $j \in U_{k'}$ with $k \neq k'$. In fact, all joint inclusions, from orders 2 to $N/2$, are $1/2$ if all members indexed are in the same partition and 0 otherwise. These second and higher order inclusion probabilities do not factor with increasing population size $N$. Thus, the number of pairwise inclusions probabilities that do not factor ($\pi_{ij} \neq 1/4$) grows at rate $\mathcal{O}(N^2)$, violating condition (A5.1).

Alternatively, we could embed the partitioning procedures within strata, where the strata are created according to rank order, have a fixed size, and the number of strata grow with population size $N$. For example grouping every 50 units into a stratum, then partitioning within each. Such a modification is relatively minor, but leads to factorization for all but $\mathcal{O}(N)$ pairwise inclusion probabilities. This can be visualized as the diagonal blocks in the full pairwise inclusion matrix (see Figure 3).

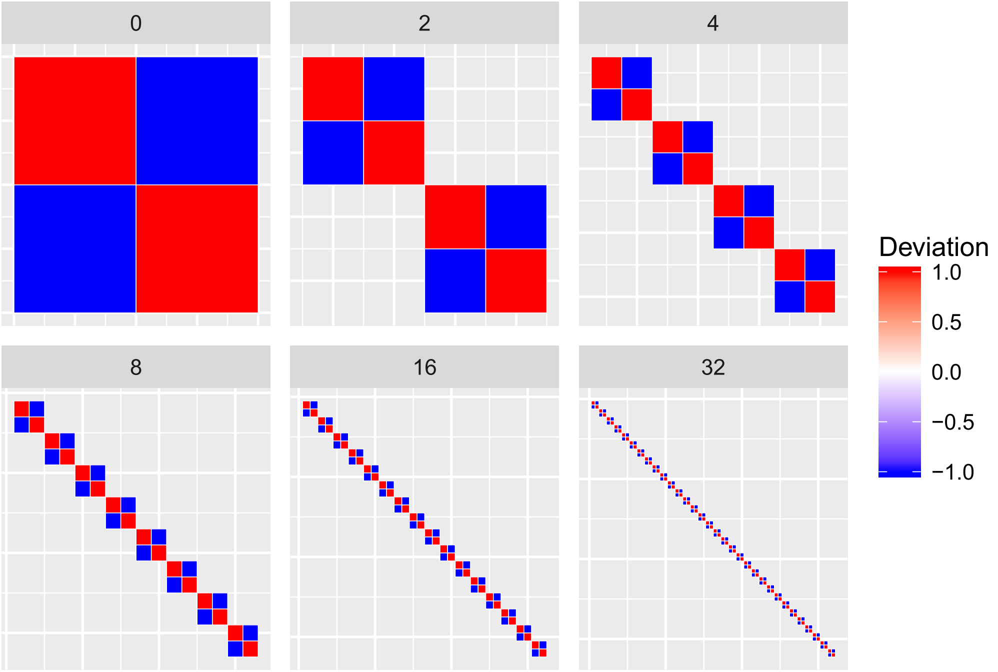

Figure 3: Matrix $\{i, j\}$ of deviations from factorization $(\pi_{ij}/(\pi_i \pi_j) - 1)$ for an equal probability dyadic partition design by number of strata (0 to 32). Empty cells correspond to 0 deviation (factorization). Created with 'ggplot2' (Wickham, 2009).

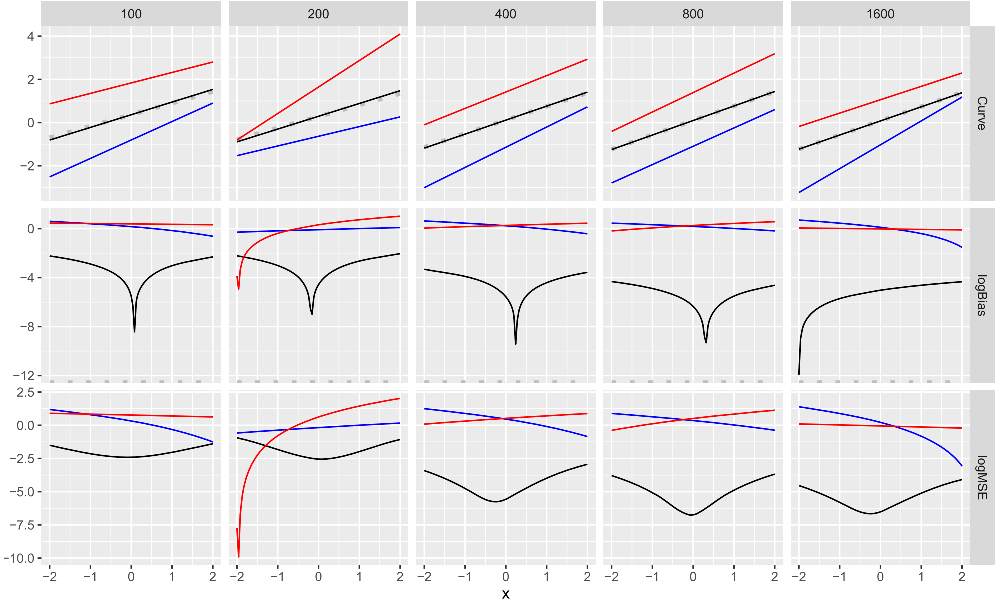

Figure 4: The marginal estimate of $\mu = f(x_1)$ under a linear model and one-stage sampling design with dyadic partitions. Compares the whole population curve (broken grey) to binary partition (red and blue/light and dark grey) and the stratified partition sample (black) - one stratum per 50 individuals, each divided into a binary partition, repeated 100 times. Top to bottom: estimated curve, log of absolute bias, log of mean square error. Left to right: doubling of population size (100 to 1600). Created with 'ggplot2' (Wickham, 2009).

For each $N \in \{100, 200, 400, 800, 1600\}$, we generate a single population and compare the relative convergence of the original dyadic partitions and the stratified versions. Figure 4 compares the bias and mean square error (MSE) of the two partitions (red and blue) compared to the average of 100 samples from the stratified version (black). It's clear that as the population size (and sample size) grows, the bias of the two partitions does not go away (the variability is due to a single realization of the population at each size), while the overall bias and MSE of the stratified version clearly decreases with increasing N, consistent with the theory.

# 5 Application to the NSDUH

A simple logistic model of current (past month) smoking status by past year major depressive episode (MDE) was fit via the survey weighted pseudo-posterior as described in Section 4 using both equal and probability-based analysis weights for 41,700 adults from the 2014 NSDUH public use data set (Figure 5). It is reasonable to assume that equal weights lead to higher estimates of smoking, as young adults are more likely to

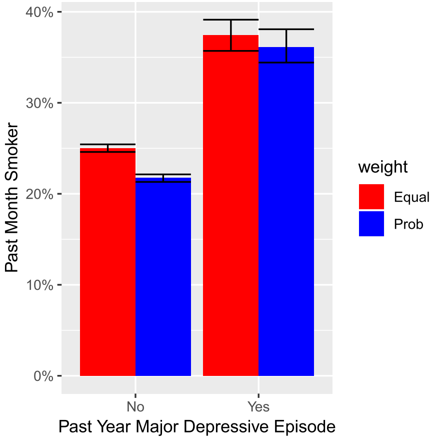

Figure 5: Posterior estimates (mean and 95% intervals) for adults in the US who smoked cigarettes in the past month by past year major depressive episode, using equal weights (red/light grey) and probability based analysis weights (blue/dark grey) based on the 2014 National Survey on Drug Use and Health. Created with 'ggplot2' (Wickham, 2009).

smoke and are over-sampled. Although the differences (after survey weighting) may seem relatively small, the effect size is policy-relevant and reflects a large number of individuals in the population.

A more dramatic difference between equal weights and survey weights can be seen when comparing the rates of new marijuana users between 33,500 youth and 21,300 adults (the at-risk respondents) (Figure 6). Because younger adults are over-sampled relative to older adults and the initiation of marijuana is rarer for older adults, the survey weighted estimated rate for adults (in the population) is nearly half that as under equal weighting. The decrease in estimated adult new marijuana users under survey weighting inflates the estimated difference between adults and youths as compared to equal weighting.

Based on the theoretical results and the simulation study presented in this paper, we have justification that the probability-based weights have removed any bias and provide consistent estimation for the NSDUH sampling design. The large number of strata and the asymptotically independent first stage of selection creates factorization for all but $\mathcal{O}(N)$ pairwise inclusion probabilities, even though the clustering and the sorting of units before selection may be informative (i.e. related to the outcome measure).

# 6 Conclusions

This work is motivated by the discrepancy between the theory available to justify consistent estimation for survey sample designs and the practice of estimation for complex, multistage cluster designs such as the NSDUH. Previous requirements for approximate

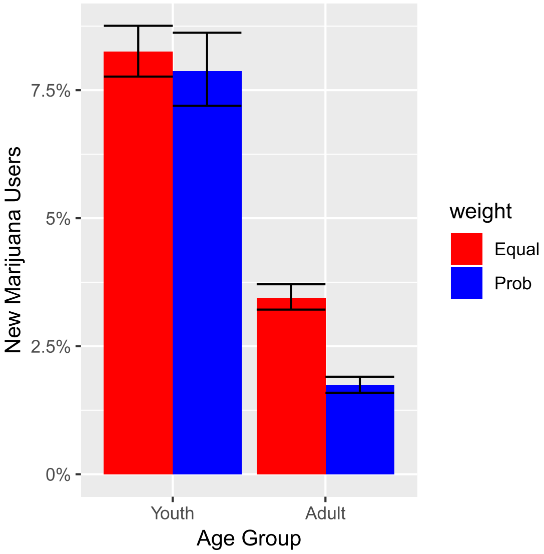

Figure 6: Posterior estimates (mean and 95% intervals) individuals who first use marijuana in the past year (2013-2014) by age group (Youth/Adult), using equal weights (red/light grey) and probability based analysis weights (blue/dark grey) based on the 2014 National Survey on Drug Use and Health. Created with ‘ggplot2’ (Wickham, 2009).

or asymptotic factorization of joint sampling probabilities exclude such designs, leaving the practitioner unable to fully justify their use. We have presented an alternative requirement that allows for unrestricted sampling dependence to persist asymptotically rather than to attenuate. For example, dependence between units within a cluster is unrestricted provided that the cluster size is bounded and dependence between clusters attenuates. This dependence can be positive (joint selection) or negative (mutual exclusion). Results are further demonstrated via a simulation study of a simplified NSDUH design.

Additional simulations expand our understanding of the impact of sorting. While the direct application of these methods can lead to dependence among all units (effectively sampling one cluster of infinite size), embedding these features within stratified or clustered designs can be justified (for subsequent estimation using marginal sampling weights) by our main results and performs well in simulation and in practice. For example, geographic units sorted along a gradient can now be fully justified for the NSDUH, because the sampling along this gradient occurs independently across a large number of strata. Similarly, the nested sampling of households within census blocks within CBSAs for the CE survey is justified due to the large number of CBSAs sampled.

With this work, the use of the sample weighted pseudo-posterior (Savitsky and Toth, 2016) is now available to a much wider variety of survey programs. We note that while establishing consistency is essential, understanding other properties of pseudo-posteriors such as credible intervals, still requires more research. Similarly, model selection for posteriors could be extended to the survey weighted pseudo-posteriors, for example the WAIC statistic (Gelman et al., 2014).

Lastly, although beyond the scope of this work, the properties of the sampling design needed for consistent estimation may also provide insight into methods which use subsampling of the complete data set (not the survey sample) as a tool for computational scalability. The results of this work provide some insight on the types of sampling designs which might be expected to work (bounded residual dependence of $\mathcal{O}(N)$) and those which would likely fail (residual dependence of order $\mathcal{O}(N^2)$). For example, an approach for scaling logistic regression for large data in Wang et al. (2018) uses subsampling of individuals without dependence ($\pi_{ij} = \pi_i \pi_j$). The main result of the current work suggests that more complex sampling such as multistage cluster sampling may also be used, if for example, a multilevel model were being fit to data with a nested structure such as patients within hospitals.

## Supplementary Material

Appendices for "Bayesian Estimation Under Informative Sampling with Unattenuated Dependence" (DOI: 10.1214/18-BA1143SUPP; .pdf). Supplementary appendices are provided online to support the proof of Theorem 1 and to demonstrate example code and MCMC diagnostics.

## References

Binder, D. A. (1983). "On the variances of asymptotically normal estimators from complex surveys." *International Statistical Review*, 51: 279–92. MR0731144. doi: `https://doi.org/10.2307/1402588`. 61

Breslow, N. E. and Wellner, J. A. (2007). "Weighted Likelihood for Semiparametric Models and Two-phase Stratified Samples, with Application to Cox Regression." *Scandinavian Journal of Statistics*, 34(1): 86–102. MR2325244. doi: `https://doi.org/10.1111/j.1467-9469.2006.00523.x`. 64

Brewer, K. (1975). "A simple procedure for $\pi$pswor." *Australian Journal of Statistics*, 17: 166–172. MR0440754. 69

Carpenter, B. (2015). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1). 68

Center for Behavioral Health Statistics and Quality (2015a). "Section 1: Adult Mental Health Tables." In *2014 National Survey on Drug Use and Health: Mental Health Detailed Tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. 60

Center for Behavioral Health Statistics and Quality (2015b). "Section 2: Tobacco Product and Alcohol Use Tables." In *2014 National Survey on Drug Use and Health:*

*Detailed Tables*. Rockville, MD: Substance Abuse and Mental Health Services Administration. 60

Chambers, R. and Skinner, C. (2003). *Analysis of Survey Data*. Wiley Series in Survey Methodology. Wiley. MR1978840. doi: https://doi.org/10.1002/0470867205. 60, 61

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and Computing*, 24(6): 997–1016. MR3253850. doi: https://doi.org/10.1007/s11222-013-9416-2. 75

Ghosal, S., Ghosh, J. K., and Vaart, A. W. V. D. (2000). "Convergence rates of posterior distributions." *The Annals of Statistics*, 28(2): 500–531. MR1790007. doi: https://doi.org/10.1214/aos/1016218228. 64

Ghosal, S. and van der Vaart, A. (2007). "Convergence rates of posterior distributions for noniid observations." *The Annals of Statistics*, 35(1): 192–223. MR2332274. doi: https://doi.org/10.1214/009053606000001172. 65, 66

Godambe, V. P. and Thompson, M. E. (1986). "Parameters of super populations and survey population: their relationship and estimation." *International Statistical Review*, 54: 37–59. MR0962931. doi: https://doi.org/10.2307/1403139. 61

Heeringa, S. G., West, B. T., and Berglund, P. A. (2010). *Applied Survey Data Analysis*. Chapman and Hall/CRC. 61

Holt, D., Smith, T. M. F., and Winter, P. D. (1980). "Regression Analysis of Data from Complex Surveys." *Journal of the Royal Statistical Society. Series A (General)*, 143(4): 474–487. MR0603746. doi: https://doi.org/10.2307/2982065. 58

Isaki, C. T. and Fuller, W. A. (1982). "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association*, 77: 89–96. MR0648029. 60

Kish, L. and Frankel, M. R. (1974). "Inference from complex samples (with discussion)." *Journal of the Royal Statistical Society, Series B*, 36: 1–37. MR0365812. 61

Morton, K. B., Aldworth, J., Hirsch, E. L., Martin, P. C., and Shook-Sa, B. E. (2016). "Section 2, Sample Design Report." In *2014 National Survey on Drug Use and Health: Methodological Resource Book*. Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. 59, 69

Pfeffermann, D., Krieger, A., and Rinott, Y. (1998). "Parametric distributions of complex survey data under informative probability sampling." *Statistica Sinica*, 8, 1087–1114 (1998). MR1666233. 60, 61

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology*, 18: 209–217. 61

Savitsky, T. D. and Toth, D. (2016). "Bayesian estimation under informative sampling." *Electronic Journal of Statistics*, 10(1): 1677–1708. MR3522657. doi: https://doi.org/10.1214/16-EJS1153. 58, 60, 61, 62, 63, 64, 66, 67, 74

Toth, D. and Eltinge, J. L. (2011). “Building consistent regression trees from complex sample data.” *Journal of the American Statistical Association*, 106(496): 1626–1636. MR2896862. doi: https://doi.org/10.1198/jasa.2011.tm10383. 60

Wang, H., Zhu, R., and Ma, P. (2018). “Optimal subsampling for large sample logistic regression.” *Journal of the American Statistical Association*, 113(522): 829–844. MR3832230. doi: https://doi.org/10.1080/01621459.2017.1292914. 75

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL http://ggplot2.org 70, 71, 72, 73, 74

Williams, M. R. and Savitsky, T. D. (2018). “Bayesian pairwise estimation under dependent informative sampling.” *Electronic Journal of Statistics*, 12(1): 1631–1661. MR3806435. doi: https://doi.org/10.1214/18-ejs1435. 61, 62

Williams, M. R. and Savitsky, T. D. (2019). “Supplementary Material for “Bayesian Estimation Under Informative Sampling with Unattenuated Dependence”.” *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1143SUPP. 66

Yi, G. Y., Rao, J. N. K., and Li, H. (2016). “A Weighted Composite Likelihood Approach for Analysis of Survey Data under Two-level Models.” *Statistica Sinica*, 26: 569–587. MR3497760. 61

## Acknowledgments

We are very grateful to the editor, associate editor and referee for their comments. The views expressed by the authors do not reflect the official position of the National Science Foundation, the U.S. Bureau of Labor Statistics, or the United States government.