

Bayesian pairwise estimation under dependent informative sampling

Matthew R. Williams

*Substance Abuse and Mental Health Services Administration
5600 Fishers Ln., Rockville, MD 20857 USA
e-mail: Matthew.Williams@samhsa.hhs.gov*

Terrance D. Savitsky

*U.S. Bureau of Labor Statistics
2 Massachusetts Ave. N.E., Washington, D.C. 20212 USA
e-mail: Savitsky.Terrance@bls.gov*

Abstract: An informative sampling design leads to the selection of units whose inclusion probabilities are correlated with the response variable of interest. Inference under the population model performed on the resulting observed sample, without adjustment, will be biased for the population generative model. One approach that produces asymptotically unbiased inference employs marginal inclusion probabilities to form sampling weights used to exponentiate each likelihood contribution of a pseudo likelihood used to form a pseudo posterior distribution. Conditions for posterior consistency restrict applicable sampling designs to those under which pairwise inclusion dependencies asymptotically limit to 0. There are many sampling designs excluded by this restriction; for example, a multi-stage design that samples individuals within households. Viewing each household as a population, the dependence among individuals does not attenuate. We propose a more targeted approach in this paper for inference focused on pairs of individuals or sampled units; for example, the substance use of one spouse in a shared household, conditioned on the substance use of the other spouse. We formulate the pseudo likelihood with weights based on pairwise or second order probabilities and demonstrate consistency, removing the requirement for asymptotic independence and replacing it with restrictions on higher order selection probabilities. Our approach provides a nearly automated estimation procedure applicable to any model specified by the data analyst. We demonstrate our method on the National Survey on Drug Use and Health.

Keywords and phrases: Survey sampling, sampling weights, quantile regression, non-linear regression, Markov chain Monte Carlo.

Received September 2017.

1. Introduction

The primary interest of the data analyst is to perform inference about a finite population generated from an unknown model, P_0 . The observed data are collected from a sample subsequently taken from that finite population under a known sampling design distribution, P_ν , that 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 the sample is different from that in the population. Savitsky & 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, $\pi_i = \mathbb{E}_{P_\nu}(\delta_i)$, for units, $i = 1, \dots, n$, where n denotes the number of units in the observed sample. The inclusion of unit, i , from the population, U , in the sample is indexed by $\delta_i \in \{0, 1\}$ and is governed by P_ν . This construction intends to account for the sampling design distribution, P_ν , to perform inference on the population generating distribution, P_0 , estimated on the observed sample taken from a finite population. Although we typically expect dependence to be induced among the sampled observations by the law for sampling, P_ν — for example, under sampling without replacement — the use of weights composed from first order inclusion probabilities ignores this dependence; hence, condition (A5) in Savitsky & Toth (2016) restricts the class of sampling designs to those where the pairwise dependencies among units attenuate to 0 in the limit of the population size, N , (at order N) to guarantee posterior consistency of the pseudo posterior distribution estimated on the sample data, at P_0 (in L_1).

While many sampling designs will meet this criterion, many won’t; for example, a two-stage clustered sampling design where the number of clusters increases with N , but the number of units in each cluster remain relatively fixed such that the dependence induced at the second stage of sampling never attenuates to 0. A common example are designs which select households as clusters.

Researchers and policy makers may be interested in the relationship between the behaviors of individuals living together (such as parents and children or spouses). This creates a sub-population of individuals defined by the behaviors of *other* members of the household, where these joint or conditional behaviors (such as substance use) are only observed through the survey. Substance use, however, is not observed in the population, but only for respondents in the sample. So the sub-population of interest is constructed by a conditioning event based on the reported substance use of other units in the sample. Sampling weights defined based on marginal inclusion probabilities are formed using quantities (e.g., size variable(s)) observed for *all units in the population* (within each stage of sampling) and aren’t designed to perform inference on a sub-population defined by information only available from *other units in the sample*.

1.1. Examples

We next outline some examples of survey programs that employ informative sampling designs under which estimation using sampling weights formed from marginal inclusion probabilities would not be guaranteed to produce a consistent result under Savitsky & Toth (2016). *Example 1:* The Current Expenditure (CE) survey is administered to U.S. households by the U.S. Bureau of Labor

Statistics (BLS) 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 multi-stage 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 or probability of purchase for a variety of goods and services. The CE sampling design is one where the number of clusters drawn increases in the limit of the population size, N , but the number of Census blocks per cluster remains relatively fixed such that we do *not* expect an attenuation of the pairwise dependencies (induced within cluster) of the secondary sampling units (Census blocks) such that the pseudo posterior formulated from marginal inclusion probabilities would not be guaranteed to achieve a consistent result under this sampling design.

Example 2: The motivating survey for the pairwise weighting method that we introduce in this paper is 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 multi-stage state-based design, with the earlier stages defined by geography within each state in order to select households (and group quarters) nested within these geographically-defined primary (or first-stage) sampling units (PSUs). Individuals or pairs of individuals are subsequently sampled from selected households. Viewing each household as a (mini) population, it is clear that the number of individuals residing in a household (of size, N_h) remains fixed in the limit of N , such that there is always unattenuated sampling dependence among those individuals.

Researchers and policy makers may be interested in the substance use of one member of a household — for example, a household that includes two spouses living together (which we term, a “spouse-spouse” household) — based on the substance use of another member of the household (e.g., their spouse), which is only observed in a subset of the sample and not in the entire sample or the population. Weights constructed on marginal inclusion probabilities may not map back to the sub-population (formed by conditioning on the self-reported behavior of the spouse) under informative sampling of the sub-population because the event (substance use by a spouse in the household) is only observed in the sample, whereas these weights are constructed only from quantities observed in the population. We illustrate this potential problem of using marginal weights for sub-population inference based on self-reported alcohol use of household members from the NSDUH.

1.2. Population model estimation

The target audience for this article are data analysts who wish to perform some distributional inference using data obtained from an informative sample

design on a population using a model they specify, P_{λ} , equipped with density, $p(x_i|\lambda)$, $\lambda \in \Lambda$, to estimate unknown, true model P_{λ_0} with density, $p(x_i|\lambda_0)$. We discuss, in the next section, how the limited literature on this topic does not adequately provide a general method for making distributional inference on a population model formulated by the data analyst while adjusting for the unequal probabilities of selection.

In this article, we propose an approach that replaces the pseudo likelihood of Savitsky & Toth (2016), $p(x_i|\delta_i = 1, \lambda)^{w_i}$, where the sampling weight, $w_i \propto 1/\pi_i$, with an approach that incorporates pairwise (second order) inclusion probabilities that provide some information about the dependence among sampled units induced by P_{ν} . The revised pseudo likelihood we will use in this paper is formed from pairwise terms, $[p(x_i|\delta_i = 1, \lambda) \times p(y_j|\delta_j = 1, \lambda)]^{w_{ij}}$, for $i, j \in U$, with $w_{ij} \propto 1/\pi_{ij}$, $\pi_{ij} = \mathbb{E}_{P_{\nu}}(\delta_i\delta_j)$. The use of weights constructed from pairwise inclusion probabilities conveys more information about the dependence induced by the joint sampling design distribution among the sampled units. Our approach retains the attractive feature of Savitsky & Toth (2016) of asymptotically unbiased inference for P_0 under any model specified by the data analyst without altering the geometry of the Markov chain Monte Carlo (MCMC) sampler. Our new approach also does not require the data analyst to have information about the sampling design, other than the (symmetric) matrix of pairwise inclusion probabilities for the lowest level units. Under many common designs, only a smaller block diagonal subset of this matrix may be needed. For example, these blocks may be households. The incorporation of second order inclusion probabilities, however, will broaden the class of sampling designs under which automated inference about P_0 may be performed by not restricting the pairwise dependence among the sampled units to attenuate to 0.

1.3. Review of methods to account for informative sampling

Research activity that incorporates sampling weights built from marginal (or first order) unit inclusion probabilities to estimate population quantities under an informative sample has surged. Recent works by Dong et al. (2014); Rao & Wu (2010); Kunihama et al. (2014); Si et al. (2015) incorporate first order sampling weights, but under a single or fixed formulation for the population generation model — typically an empirical likelihood or Dirichlet process mixture for flexibility — with a focus on performing inference about simple population statistics, such as the total and mean. These approaches focus on design inference, rather than inference from a model of interest specified by the data analyst, the latter of which is our focus in this paper. Savitsky & Toth (2016), alternatively, formulate a pseudo posterior distribution as a plug-in estimator, using first order sampling weights to allow the data analyst to perform inference from any population generating model that they specify.

As earlier mentioned, Savitsky & Toth (2016) define conditions that restrict allowable sampling designs such that frequentist consistency of their pseudo posterior approximation is guaranteed. One of these conditions requires sampling designs where the pairwise sample inclusion dependencies among units

attenuates to 0 in the limit of the population size. While they discuss many sampling designs that satisfy this restriction, many do not — such as the CE and NSDUH examples we earlier discussed. In a similar fashion to this paper, Yi et al. (2016) start with a pairwise likelihood construction that incorporates sampling weights based on second order (pairwise) inclusion probabilities to perform (pseudo) maximum likelihood estimation in order to capture second order dependence among sampled units. However, they construct the population generating model to explicitly match the sampling design, which they restrict to a 2-stage sampling design. So like the recent works using sampling weights composed from first order inclusion probabilities, they require a specific formulation of the population model that does not allow the data analyst to perform inference on a model of their choosing. By contrast, our approach for incorporating sampling weights based on pairwise inclusion probabilities allows model inference under *a large class of* population generating models specified by the data analyst. We do not require the use of a 2-stage sampling design or even that the population model and sampling designs match; rather, in the sequel we will formulate conditions that, together, define a class of sampling designs under which frequentist consistency of our (improper) pseudo posterior approximation is guaranteed. Our use of second order (pairwise) weights allows us to broaden the class of allowable samplings designs under which our pseudo posterior estimator contracts on the true population generating distribution by eliminating the requirement for pairwise dependencies to attenuate to 0.

We introduce the concept of pairwise weighting for a practical implementation in Section 2 for a generic parametric population model. A rigorous theoretical exposition enumerates conditions under which the L_1 consistency of our weighted estimator composed from joint inclusion probabilities is guaranteed is presented in Section 3. This section may be skipped for non-interested readers without loss of continuity. The construction for the population model designed to analyze the conditional quantiles for alcohol consumption of individuals as a function of age that we will use for both a simulation study and applications follows in Section 4. A simulation study is conducted in Section 5, where we assess the relative bias of marginal and pairwise weighting schema for estimation of the conditional quantiles of alcohol consumption for an individual living in a spouse-spouse pair, conditioned on the alcohol consumption of their partner. An application is made to the National Survey on Drug Use and Health (NSDUH) in Section 6. A concluding discussion follows in Section 7.

2. Pairwise weighting to account for informative sampling

We begin by constructing the pseudo likelihood and associated pseudo posterior density under any analyst-specified prior formulation on the model, $\boldsymbol{\lambda} \in \Lambda$.

Suppose there exists a Lebesgue measurable population-generating density, $p(x|\boldsymbol{\lambda})$, indexed by parameters, $\boldsymbol{\lambda} \in \Lambda$. Let $\delta_i \in \{0, 1\}$ denote the sample inclusion indicator for units $i = 1, \dots, N$ from the population under sampling without replacement. The density for the observed sample is denoted by, $p(x_o|\boldsymbol{\lambda}) = p(x|\delta_i = 1, \boldsymbol{\lambda})$, where “o” indicates “observed”.

The following plug-in estimator for the posterior density incorporates sampling weights formulated from pairwise inclusion probabilities under the analyst-specified model for $\boldsymbol{\lambda} \in \Lambda$,

$$p^\pi(\boldsymbol{\lambda}|\mathbf{x}_o, \mathbf{w}) \propto \left[\prod_{i,j=1}^n \{p(x_{o,i}|\boldsymbol{\lambda})p(x_{o,j}|\boldsymbol{\lambda})\}^{w_{ij}} \right] p(\boldsymbol{\lambda}) \quad (1)$$

$$\propto \left[\prod_{i=1}^n \prod_{j \neq i \in S} p(x_{o,i}|\boldsymbol{\lambda})^{w_{ij}} \right] p(\boldsymbol{\lambda}) \quad (2)$$

$$= \left[\prod_{i=1}^n p(x_{o,i}|\boldsymbol{\lambda})^{\sum_{j \neq i} w_{ij}} \right] p(\boldsymbol{\lambda}) \quad (3)$$

$$= \left[\prod_{i=1}^n p(x_{o,i}|\boldsymbol{\lambda})^{w_i^*} \right] p(\boldsymbol{\lambda}) \quad (4)$$

where we have used the independence of the (x_i) , conditioned on $\boldsymbol{\lambda}$, under P_{λ_o} , to rearrange terms in the product to achieve Equation 4, which exponentiates the likelihood contribution of unit i by the sum of the sampling weights, $\{w_{ij} \propto 1/\pi_{ij}\}$, formulated to be inversely proportional to *pairwise* or second order inclusion probabilities. The collection of pairwise inclusion probabilities that, together, are used to formulate, w_i^* , the sampling weighted exponent for unit i , represent all pairs by which unit i enters the observed sample. The sum of the pairwise sampling weights for each unit, i , assigns the relative importance of the likelihood contribution for that observation to approximate the likelihood for the population. We use p^π to denote the noisy approximation to distribution, p , and we make note that the approximation is based on the data, \mathbf{x}_o , and sampling weights, $\{\mathbf{w}^*\}$, confined to those units *included* in the realized sample, $\{i \in U : \delta_i = 1\}$, where U denotes a population of units indexed by $i = 1, \dots, N$. We drop the subscript, “o”, in the sequel when referring to the observations and replace it with a conditioning statement.

The total estimated posterior variance is regulated by the sum of the sampling weights to approximate the amount of information in our observed sample. We define unnormalized second order weights, $\{w_{ij} = 1/\pi_{ij}\}$. The agency which conducts the survey would construct, $\tilde{w}_i = \sum_{j \neq i} w_{ij} \delta_i \delta_j / (N - 1)$ for each unit $i \in (1, \dots, N)$ in the population. The sum of the pairwise weights is divided by $N - 1$ to account for the $N - 1$ times that unit i appears in the combination of pairs in the population (each of which produces a likelihood contribution), which adjusts for duplicative likelihood contributions in the population. Due to the presence of the joint indicator $\delta_i \delta_j$, the summation can be taken over the second order weights in the *observed* sample rather than the entire population. The data analyst will see the weights for *observed* units published in the sample. They will need to normalize the (\tilde{w}_i) by computing $w_i^* = \frac{n}{\sum_{i=1}^n \tilde{w}_i} \tilde{w}_i$, $i = 1, \dots, n$ to sum to the sample size, n . This normalization step approximates the amount

of information in our sample of size, n . It regulates the amount of posterior uncertainty estimated by application of the population model under use of the (w_i^*) to the observed sample.

The weight for an individual is formulated by a sum of the components constructed from the multiplication of (inverse) joint probabilities across the sampling stages. We demonstrate in Appendix B.3 that under a multi-stage sampling design, such as the sampling of individuals within households within geographic segments for our NSDUH example, this sum is dominated by terms that essentially factor due to nearly independent sampling in the earlier stages. The result is that the pairwise-formulated weight, w_i^* , quickly converges (under increasing sample size) to the sampling weight formed from marginal inclusion probabilities, w_i . Therefore, individual weights formed from the sum of inverse joint inclusion probabilities won't provide any improvement over marginal weights for inference on the population of individuals, which corresponds to all possible pairs of individuals in the population. However, without loss of generality, we can focus inference on a sub-population of pairs, such as the population of co-habiting spouses. Then we formulate w_i^* to include only a single pairwise term, $w_{ij|\ell}$, in the case a single pair is sampled within a household, ℓ , and individuals j and i are co-sampled. Our reformulated pairwise weighting scheme treats spouse pairs (which are nested within households) as a population of interest. It will be shown to reduce bias relative to the use of weights formulated from marginal (or individual) inclusion probabilities for inference about the behavior of one member of a pair conditioned on the behavior of the other under an informative selection of the conditioning event.

3. Pairwise pseudo posterior consistency

3.1. Pseudo posterior distribution

We next introduce a formal exposition of the pairwise-weighted pseudo posterior distribution and construct conditions under which it is guaranteed to contract on the true generating distribution. Our asymptotics are driven by $\nu \in \mathbb{Z}^+$, which indexes a sequence of infinite populations, $\{U_\nu\}$, each of size N_ν , where $N_{\nu'} > N_\nu$ for $\nu' > \nu$, such that the finite population size grows with ν . In each increment of ν , we identify units in a finite population, generate random variables, \mathbf{X}_ν from population generating distribution, P_0 , from which we, subsequently, take a sample of size, n_ν , governed by sampling design distribution, P_ν . A sampling design is defined by placing a *known* distribution on a vector of inclusion indicators, $\boldsymbol{\delta}_\nu = (\delta_{\nu 1}, \dots, \delta_{\nu N_\nu})$, linked to the units comprising the population, U_ν . Choice of design (i.e. specification of the distribution for $\boldsymbol{\delta}_\nu$) may depend on values from the population which is generated from a hypothetical distribution, P_0 , independently, but not identically (*inid*), of which independent and identical (*iid*) is a special case. The sampling distribution is subsequently used to take an *observed* random sample of size $n_\nu \leq N_\nu$. We make no parametric or distributional assumption for the true generating model,

P_0 , or its estimate, P , in the sequel. So we do not assume that P_0 is indexed by parameters, λ , in the form of P_{λ_0} . Our conditions needed for the main result, to follow, employ known second-order or pairwise unit inclusion probabilities, $\pi_{\nu ij} = \mathbb{E}_{P_\nu} \{\delta_{\nu i} \delta_{\nu j}\}$ for all $i \neq j \in U_\nu$, rather than the marginal inclusion probabilities, $\pi_{\nu i} = \mathbb{E}_{P_\nu} \{\delta_{\nu i}\}$ for $i \in U_\nu$ used in Savitsky & Toth (2016), which are both obtained from the joint distribution over $(\delta_{\nu 1}, \dots, \delta_{\nu N_\nu})$. The dependence among unit inclusions in the sample contrasts with the usual *iid* draws from P_0 . We denote the sampling distribution by P_ν .

Under informative sampling, the inclusion probabilities (typically marginal) are formulated to depend on the finite population data values, $\mathbf{X}_{N_\nu} = (\mathbf{X}_1, \dots, \mathbf{X}_{N_\nu}) \stackrel{\text{ind}}{\sim} P_0$. Since the resulting balance of information (where information describes the joint likelihood) would be different in the sample, a posterior distribution from $(\mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu})$ that ignores the distribution for δ_ν will not lead to consistent estimation. In addition, under a complex sampling design with multiple stages, correlations are typically induced among the inclusions for some or all units.

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,

$$p^\pi(\mathbf{X}_i \delta_{\nu i}) := p(\mathbf{X}_i) \frac{1}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}}, \quad (5)$$

that weights each density contribution, $p(\mathbf{X}_i)$, by the sum of all of its inverse pairwise inclusion probabilities, which together represent all pairwise paths by which unit i may enter a selected sample. The employment of pairwise inclusion probabilities partially accounts for the dependence of among unit inclusions induced by P_ν . The sum of terms for each i is divided by $N_\nu - 1$ because each individual is present in $N_\nu - 1$ population pair terms in the summation, each of which has expectation with respect to P_ν equal to 1. So the normalization of the summation term ensures that the expectation of the logarithm of the density with respect to P_ν is unbiased. Our construction re-weights the likelihood contributions defined on those units randomly-selected for inclusion in the observed sample ($\{i \in U_\nu : \delta_{\nu i} = 1\}$) to *approximate* the balance of information in U_ν , from which we construct the associated pseudo posterior,

$$\Pi^\pi(B | \mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu}) = \frac{\int_{P \in B} \prod_{i=1}^{N_\nu} \frac{p^\pi(\mathbf{X}_i \delta_{\nu i})}{p_0^\pi} d\Pi(P)}{\int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi(\mathbf{X}_i \delta_{\nu i})}{p_0^\pi} d\Pi(P)}, \quad (6)$$

that we use to achieve our required conditions for the rate of contraction of the pseudo posterior distribution on P_0 . $\Pi^\pi(B | \mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu})$ denotes the pseudo posterior mass placed on parameter set B contained in space, \mathcal{P} of probability measures, P . We note that both P and δ_ν are random variables defined on the space of measures (\mathcal{P} and $B \subseteq \mathcal{P}$) and possible samples, respectively. We divide the top and bottom of Equation 6 by $p_0^\pi(\mathbf{X}_i \delta_{\nu i})$, which is defined analogously as Equation 5, replacing p with p_0 . An important condition

on P_ν formulated in Savitsky & 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 show in the sequel that our use of pairwise inclusion probabilities to formulate sampling weights in the pseudo posterior distribution replaces their condition that requires marginal factorization of the pairwise inclusion probabilities with two conditions that require pairwise factorization of *third* and *fourth* order 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, π , to denote the dependence on the known sampling probabilities, $\{\pi_{\nu ij}\}_{i,j \in U_\nu}$; for example,

$$\Pi^\pi (B|\mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu}) := \Pi \left(B \middle| (\mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu}), \left(\sum_{k \neq 1 \in U_\nu} \pi_{\nu 1k}, \dots, \sum_{k \neq N_\nu \in U_\nu} \pi_{\nu N_\nu k} \right) \right).$$

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 $M > 0$.

3.2. 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 & Wellner (2007) and incorporates inverse inclusion pairwise probability weights, $\{1/\pi_{\nu ij}\}_{i,j \in U_\nu}$, to account for the informative sampling design,

$$\mathbb{P}_{N_\nu}^\pi = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{1}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} \delta(\mathbf{X}_i), \tag{7}$$

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, $\mathbb{P}_{N_\nu} = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \delta(\mathbf{X}_i)$, used to approximate $P \in \mathcal{P}$, the distribution hypothesized to generate the finite population, U_ν .

We follow the notational convention of Ghosal et al. (2000) and define the associated expectation functionals with respect to these empirical distributions by

$\mathbb{P}_{N_\nu}^\pi f = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{1}{(N_\nu-1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} f(\mathbf{X}_i)$. 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, $\mathbb{G}_{N_\nu}^\pi = \sqrt{N_\nu} (\mathbb{P}_{N_\nu}^\pi - P_0)$ and $\mathbb{G}_{N_\nu} = \sqrt{N_\nu} (\mathbb{P}_{N_\nu} - P_0)$.

The sampling-weighted, (average) pseudo Hellinger distance between distributions, $P_1, P_2 \in \mathcal{P}$,

$$d_{N_\nu}^{\pi,2}(p_1, p_2) = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{1}{(N_\nu-1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i)), \quad (8)$$

where $d(p_1, p_2) = \left[\int (\sqrt{p_1} - \sqrt{p_2})^2 d\mu \right]^{\frac{1}{2}}$ (for dominating measure, μ). We need this empirical average distance metric because the observed (sample) data drawn from the finite population under P_ν are no longer independent. The implication is that our consistency result applies to finite populations generated as *inid* from which informative samples are taken. The associated non-sampling Hellinger distance is specified with, $d_{N_\nu}^2(p_1, p_2) = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i))$.

3.3. Main result

We proceed to construct associated conditions and a theorem that contain our main result on the consistency of the pairwise pseudo posterior distribution under a class of informative sampling designs at the true generating distribution, P_0 . Our approach extends the main in-probability convergence result of Ghosal & van der Vaart (2007) by adding new conditions that restrict the distribution of the informative sampling design. 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)

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

(A2) (Size of space)

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

(A3) (Prior mass covering the truth)

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

(A4) (Non-zero Pairwise Inclusion Probabilities)

$$\sup_\nu \left[\frac{1}{\min_{i,k:k \neq i \in U_\nu} |\pi_{\nu ik}|} \right] \leq \gamma \geq 1, \text{ with } P_0\text{-probability 1.}$$

(A5) (Bounded Ratio of Third to Second Order Inclusion Probabilities)

$$\begin{aligned} & \sup_{\nu} \max_{i,k,\ell:k \neq \ell \neq i \in U_{\nu}} \left| \frac{\pi_{\nu i k \ell}}{\pi_{\nu i k} \pi_{\nu i \ell}} \right| \\ &= \sup_{\nu} \max_{i,k,\ell:k \neq \ell \neq i \in U_{\nu}} \left| \frac{\pi_{\nu k \ell | i}}{\pi_{\nu k | i} \pi_{\nu \ell | i} \pi_i} \right| \leq C_5, \text{ with } P_0\text{-probability 1,} \\ \text{where} \end{aligned}$$

$$\pi_{\nu k | i} = \mathbb{E}_{P_{\nu}} (\delta_{\nu k} | \delta_{\nu i} = 1), \quad \pi_{\nu k \ell | i} = \mathbb{E}_{P_{\nu}} (\delta_{\nu k} \delta_{\nu \ell} | \delta_{\nu i} = 1).$$

(A6) (Asymptotic Factorization of Fourth Order Inclusion Probabilities)

$$\begin{aligned} & \limsup_{\nu \uparrow \infty} \max_{i,j,k,\ell:i \neq j, k \neq i, \ell \neq j \in U_{\nu}} \left| \frac{\pi_{\nu i k j \ell}}{\pi_{\nu i k} \pi_{\nu j \ell}} - 1 \right| \\ &= \mathcal{O}(N_{\nu}^{-1}), \text{ with } P_0\text{-probability 1} \end{aligned}$$

such that for some constant, $C_4 > 0$,

$$N_{\nu} \sup_{\nu} \max_{i,j,k,\ell:i \neq j, k \neq i, \ell \neq j \in U_{\nu}} \left| \frac{\pi_{\nu i k j \ell}}{\pi_{\nu i k} \pi_{\nu j \ell}} - 1 \right| \leq C_4, \text{ for } N_{\nu} \text{ sufficiently large.}$$

(A7) (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 for Savitsky & Toth (2016) and restrict the growth rate of the model space (e.g., of parameters) and requires prior mass to be placed on a region containing the true value. The growth rate of the model space is captured by the logarithm of the covering number $N(\cdot)$ in Condition (A1), defined as the *minimum* number of balls of radius $\xi/36$ need to cover $\{P \in \mathcal{P}_{N_{\nu}} : d_{N_{\nu}}(P, P_0) < \xi\}$ under distance metric, $d_{N_{\nu}}$. Condition (A2) restricts the prior mass placed on the uncountable portion of the model space, \mathbf{P} , that does not include the countable space, $\mathcal{P}_{N_{\nu}}$, denoted by $\mathcal{P} \setminus \mathcal{P}_{N_{\nu}}$. The next four new conditions impose restrictions on the sampling design and associated known distribution, P_{ν} , which are similar than those specified in Savitsky & Toth (2016), but allow for a wider class of sampling designs under which consistency of the pseudo posterior formulation of Equation 5 is guaranteed by replacing the asymptotic attenuation of pairwise inclusion dependencies with restrictions on third and fourth order inclusion dependencies. Condition (A4) requires the sampling design to assign a positive probability for pairwise inclusion for every pair of units, $i, j \in U_{\nu}$. Since the maximum pairwise inclusion probability is 1, the bound, $\gamma \geq 1$. This condition is no more restrictive than the analogous condition A4 in Savitsky & Toth (2016), which bounds marginal inclusion probabilities away from 0, in the case that $\text{Cov}(\delta_{\nu i}, \delta_{\nu j}) > 0$, which implies that $\min\{\pi_{\nu i}, \pi_{\nu j}\} \geq \pi_{\nu ij} > \pi_{\nu i} \pi_{\nu j}$; otherwise, for designs where

$\text{Cov}(\delta_{\nu i}, \delta_{\nu j}) < 0$, condition (A4) is more restrictive because $\{\pi_{\nu i}, \pi_{\nu j}\} > 0$ does *not* imply $|\pi_{\nu ij}| > 0$. All pairs of units must be assigned non-zero pairwise inclusion probabilities. We make note that other than this restriction bounding pairwise inclusion probabilities away from 0, there is no required attenuation of pairwise dependencies as there is in Savitsky & Toth (2016). Instead, we add the new condition (A5) that restricts sampling designs under which the ratio of third order inclusion probabilities to the product of second order inclusion probabilities is absolutely bounded from above. This ratio approaches the condition of bounding first order inclusion probabilities away from 0 in the case that the *conditional* pairwise inclusion probabilities asymptotically factor (though such is not required). Condition (A6) requires fourth order inclusion probabilities to factor to pairwise probabilities as $N_\nu \uparrow \infty$. We note the presence of pairwise inclusion probabilities in the denominator for each our conditions (A5) and (A6), as contrasted with marginal inclusion probabilities in the analogous condition A5 in Savitsky & Toth (2016) (which requires asymptotic factorization of pairwise inclusion probabilities). The conditions of Savitsky & Toth (2016) may be viewed as requiring sampling designs that limit to the equivalent to the independent sampling of individual units, while our conditions asymptotically require designs to limit to the independent sampling of distinct *pairs* of individuals; that is, the joint inclusion probabilities of pairs, (i, j) and (k, ℓ) limit to independence in Condition (A6). Condition (A5) requires asymptotic factorization of pairwise inclusion probabilities, but only when the pairwise inclusion probability is formed by conditioning on the inclusion of a third unit, $i \in U_\nu$, which is less restrictive than the required unconditional factorization under use of marginal weights; for example, conditioned on the inclusion of individual, i , in a household, the inclusions of units j and k in that household would be asymptotically independent. Condition (A7) ensures that the observed sample size, n_ν , limits to ∞ along with the size of the partially-observed finite population, N_ν , such that the variation of information about the population expressed in realized samples is controlled.

Theorem 3.1. *Suppose conditions (A1)-(A7) hold. Then for sets $\mathcal{P}_{N_\nu} \subset \mathcal{P}$, constants, $K > 0$, and M sufficiently large,*

$$\mathbb{E}_{P_0, P_\nu} \Pi^\pi \left(P : d_{N_\nu}^\pi(P, P_0) \geq M \xi_{N_\nu} | \mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu} \right) \leq \frac{16\gamma^2 [\gamma + C_3]}{(Kf + 1 - 2\gamma)^2 N_\nu \xi_{N_\nu}^2} + 5\gamma \exp\left(-\frac{Kn_\nu \xi_{N_\nu}^2}{2\gamma}\right), \quad (9)$$

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

Proof. The proof follows exactly that in Savitsky & 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 replace their condition (A4), which bounds the inverse of marginal inclusion probabilities, with

our condition (A4), that now bounds the inverse of pairwise inclusion probabilities. We reformulate two enabling lemmas of Savitsky & Toth (2016), which we present in Appendix A, where the reliance on (their) condition (A5) requiring asymptotic factoring of pairwise unit inclusion probabilities is here replaced by conditions (A5) and (A6) that require asymptotic pairwise factoring of fourth order inclusion probabilities and boundedness in the ratio of third-to-second order inclusion probabilities. \square

We note that the rate of convergence is decreased for a sampling distribution, P_ν , that expresses a large variance in unit pairwise inclusion probabilities such that γ will be relatively larger. Samples drawn under a design that expresses a large variability in the second order sampling weights will express more dispersion in their information relative to a simple random sample of the underlying finite population. We construct $C_3 = C_4 + C_5 + 1$, such that to the extent that the third and fourth order dependencies attenuate faster than the pairwise inclusion probabilities under the pseudo posterior constructed from first order sampling weights, then the rate of contraction will be faster under our formation than in Savitsky & Toth (2016). In general, however, one would not necessarily expect a more rapid contraction under our employment of second order inclusion probabilities to form our sampling weights because the rate in both Savitsky & Toth (2016) and here is nearly optimal, as we may observe by plugging in for the rate, $\xi_{N_\nu} = \log n_\nu / \sqrt{n_\nu}$ — the optimal convergence rate reduced by a log factor — and noting that the bound in Equation 9 limits to 0. The main benefit of our approach is that it is expected to broaden the class of sampling designs (relative to Savitsky & Toth (2016)) under which the associated pseudo posterior distribution achieves a frequentist consistency result.

4. Example population model

We construct an example of a possible population model which an analyst might use to address the inferential interest of assessing the functional form of the relationship between frequency of alcohol consumption and age at conditional quantiles of interest for the population distribution of the U.S., as estimated from the 2014 National Survey on Drug Use and Health (NSDUH). This model is but one specific choice which satisfies the conditions (A1)-(A3).

We follow Reed & Yu (2009) and formulate a likelihood for each observation using the asymmetric Laplace (AL) distribution,

$$x_i \mid \mu_i, \tau, q \stackrel{\text{ind}}{\sim} \mathcal{AL}(\mu_i, \tau, q), \quad i = 1, \dots, N \quad (10)$$

where τ is a precision parameter and $q \in (0, 1)$ is the quantile of interest. We recall the AL density for observed response, x ,

$$p(x \mid \mu, \tau, q) = \tau q (1 - q) \exp(-\tau \rho_q(x - \mu)), \quad (11)$$

where

$$\rho_q(u) := \begin{cases} q|u|, & \text{if } u \geq 0 \\ (1 - q)|u|, & \text{if } u < 0 \end{cases} \quad (12)$$

To accommodate expected non-linearity in the relationship of age with the distribution for alcohol consumption, we encode age, z_1 , in a B-spline basis term,

$$\boldsymbol{\mu} = \mathbf{B}(z_1)\boldsymbol{\theta} \quad (13)$$

for $N \times (d + k)$ B-spline basis matrix, $\mathbf{B}(z_1)$, that we extend as in Clifford & Choy (2012), to convert the B-spline to a penalized (P-) spline of order k with employment of a penalty matrix, $\mathbf{Q} = \mathbf{D}'\mathbf{D}$, where \mathbf{D} has $d + k$ columns for a B-spline basis with d knots and is the discretized k^{th} difference operator. Higher values for k enforce greater smoothness restrictions in each B-spline piecewise basis (column of \mathbf{B}) under the following multivariate Gaussian prior for $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta} \mid \kappa) \propto \exp\left(-\frac{\kappa}{2}\boldsymbol{\theta}'\mathbf{Q}\boldsymbol{\theta}\right), \quad (14)$$

where $(d + k) \times (d + k)$ penalty (precision) matrix, \mathbf{Q} is of rank, d , in a similar fashion as the intrinsic conditional autoregressive prior (Rue & Held, 2005). Parameter, κ , is the smoothing, penalty parameter on which we impose a further $\mathcal{G}(1, 1)$ prior, specified with small hyperparameter settings easily overwhelmed by the data. We choose $d = 10$ and $k = 3$ such that each spline basis lies in the space of piecewise C^3 functions. Precision parameter, τ , from Equation 10 also receives a $\mathcal{G}(1, 1)$ prior.

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

$$\log \left[\prod_{i=1}^n p(x_i \mid \mu_i, \tau, q)^{w_i^*} \right] = \sum_{i=1}^n w_i^* \log p(x_i \mid \mu_i, \tau, q) \quad (15)$$

$$\begin{aligned} &= w_{\text{TOT}}^* [\log \tau + \log q + \log(1 - q)] \\ &\quad - \tau \sum_{i=1}^n w_i^* \rho_q(x_i - \mu_i), \end{aligned} \quad (16)$$

where $w_{\text{TOT}}^* = \sum_{i=1}^n w_i^*$, with sampling weights, w_i^* , as defined using joint inclusion probabilities for unit i in Section 3.1 or, alternatively, using marginal inclusion probabilities as in Savitsky & Toth (2016), to support our comparison of alternative weighting schema. We recall that we have normalized the sum of the weights such that $w_{\text{TOT}}^* = n$. Finally, we estimate the joint posterior distribution using Equation 15, coupled with our prior distributions assignments, using the NUTS Hamiltonian Monte Carlo algorithm implemented in Stan (Carpenter, 2015).

5. Simulation study

5.1. Scenarios

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 multi-stage sampling designs where the number of last stage units does not grow with overall population size). We focus our inference on the case of analyzing (some conditional quantile for) alcohol usage for a sub-population that is formed by conditioning the inclusion of a sampled individual in a spouse-spouse household based on the self-reported frequency of alcohol usage by their spouse. We construct two scenarios, where each targets a sub-population, under which we will compare the estimation performances (through bias and mean square error) of marginal versus pairwise weighting schema. These scenarios will be used for both our simulation study and following application to NSDUH:

- (S1) A sub-population target for inference is defined by those individuals who reside in a particular household configuration. So we only include the sub-sample of individuals who reside in a spouse-spouse pair for model estimation (regardless of whether their spouse is also included in the sample). This sub-population is formed using information observed in the (household) population; e.g., the household roster provides information on whether someone resides with a spouse. Inference is limited to only the responding spouse without regard to the presence of the other spouse in the sample or his or her corresponding behaviors. This scenario, in which we expect marginal inference about spouse pair members to be correct when using marginal weights, serves as a contrast to the next scenario.
- (S2) The sub-population of interest is further restricted to those spouse-spouse households where one spouse consumes alcohol above (and/or below) some threshold level frequency. This sub-population is defined based on a conditioning event (the level of alcohol consumption by one member of a spouse-spouse pair) and the condition is *not* observed in the (household) population. By construction, the conditioning event is restricted to be observed *only* in the pair sample. In contrast to the previous scenario, we expect this conditional inference between pairs to be biased when using marginal weights, but correct when using pairwise weights. The latter compensate for the differential sub-selection of the joint spouse-spouse sample from among the much larger sample of individuals who reported living with a spouse but whose spouse may not be in the sample.

5.2. Conditional populations

We consider a population of households (HHs) which each contain 3 individuals (P1, P2, P3). We assume the first two members (P1, P2) have a special relationship (such as spouse-spouse) with each other but not with P3. The response x is drawn from an AL distribution with $q = 0.5$, corresponding to the median. We let μ depend on two predictors z_1 and z_2 . The variable z_1 represents the observed information available for analysis, whereas z_2 represents information available for sampling, which is either ignored or not available for analysis. For both P1 and P3, the distributions for z_1 and z_2 are normal and exponential

distributions, respectively.

The conditional quantile distributions for P1 and P3 are identical, $\mu = f(z_1, z_2)$. The distribution for the size variable z_2 for P2 is conditional on the size variable for P1. In other words, the predictor information and the probability of selection for P2 is related to corresponding information for their spouse P1. Furthermore, the shape of the curve $f(z_1, z_2)$ is differentially defined when the size variable for P1 is above the population median. The relationship between P2's alcohol consumption and covariates has a different shape depending on whether their spouse (P1) has covariate values above a threshold (More detail is provided in Appendix B.1).

The sub-populations of P2-P1 and P2-P3 pairs within a HH each have distinct distributions for P2 for both the outcome and the joint selection probability. This is true even though the marginal distributions for outcomes and predictors are the same for their pair partners P1 and P3. In other words, we are now more likely to sample P2 within 'spouse' pairs (P1-P2) than within 'non-spouse' pairs (P2-P3), and the response distribution for P2 is skewed larger (e.g., for alcohol consumption) for 'spouse' pairs in the sample compared to the marginal population for P2.

Even though the population response x was generated from $\mu = f(z_1, z_2)$, we, subsequently, estimate the marginal model, $\mu = f(z_1)$, on the generated population and use the estimated μ from this marginal model to compare the estimated results on our observed sample (taken from the population) under the three weighting scenarios. The generation of x from size variable, z_2 , ensures that our sampling design instantiates informativeness, since inclusion probability, π , is set to be proportional to z_2 . Excluding z_2 in the target population model for x represents the lack of availability of z_2 to the data analyst and that it is a nuisance variable to the data analyst because their inferential focus is on the relationship between x and z_1 . In summary, we estimate the models under both scenarios under our informative sampling design and compare them to the population fitted model, $\mu = f(z_1)$.

Rather than sampling households and individuals directly, we wish to more closely mimic multi-stage cluster designs such as the NSDUH. Sampling first occurs for groups of households within primary sampling units (PSUs). Then households (HHs) and pairs within households can be selected. Unequal probabilities of selection are possible at each stage by aggregating the size measure up to the sampling unit (PSU or HH). See Appendix B.2 for more details.

5.3. Calculation of weights

From the three stages of sampling (PSUs, HHs, pairs) there are four weight components available to use:

1. PSU: $w_1^k = 1/\pi_k$, the inverse of the probability of selecting PSU k .
2. HH: $w_2^{j|k} = 1/\pi_{j|k}$, the inverse of the conditional probability of selecting HH j given PSU k has been selected.

3. Individual: $w_3^{i|jk} = 1/\pi_{i|jk}$, the inverse of the conditional probability of selecting individual i given HH j and PSU k are selected.
4. Pairwise: $w_3^{i,i'|jk} = 1/\pi_{i,i'|jk}$, the inverse of the joint probability of selecting individuals i and i' as a pair in HH j given the household and PSU k are selected.

In general, each stage could have two sets of weights from both first and second order components. For this example, the joint probabilities of selection within each of the first two stages are considered negligible.

Based on these four weight components, the first order or marginal weight is simply the inverse of the probability of selecting an individual: $w_i^{(1)} = w_1^k w_2^{j|k} w_3^{i|jk}$. For second order weights, we set the HH as the unit of analysis and construct each pairwise weight *within* HH for individual, i : $w_i^{(2p)} = w_1^k w_2^{j|k} w_3^{i,i'|jk} / (N_{p_j} - 1)$, where i' is the co-sampled individual in HH, j , that includes units, (i, i') . We normalize by the number of pairs in which i is a member in the domain of interest within each household, $(N_{p_j} - 1)$, because each roster of the HH is treated as a population (and the entire population is constructed as the collection of household populations). The distribution of household pairwise weights is similar to the marginal weights, but the former have a slightly heavier tail. While the overall distributions are similar, the ratio of the marginal weights and the pairwise weights reveals that the pairwise weights do in fact redistribute the marginal weight, increasing the weight of some individuals and decreasing the weight for others. More discussion on the formulation of the household pairwise weights and some example properties of these different weights are demonstrated in Appendix sections B.3 and B.4.

5.4. Results

For simplicity and scalability to small sample sizes, we model both the population and sample using $d = 5$ knots and polynomials of degree $k = 2$. Each column of Figure 1 displays the fitted curves, bias and mean square error (MSE) for scenario (S1) that includes the full sample of P2's (anyone living with a spouse) for a particular average sample size. Both weighting methods remove the bias compared to the equal weighting. The first order or marginal weights show a slight, but persistent edge in MSE likely due to less variability in the marginal weights.

Figures 2 presents results under scenario (S2), where the P2 sub-population of interest is *conditioned* on whether the observed response, x , for P1 (the other spouse) is above 10, a value which is close to the median of x (Results for when P1 is below 10 are similar and therefore not included). The household pairwise weights remove more bias and lead to smaller MSE than do the marginal weights because the conditioning event on the x value for P1 is only observed in the P1-P2 (joint spouse) sub-sample, but not in the full P2 sample or the (household) population.

Therefore, the computed value of each household pairwise weight is able to

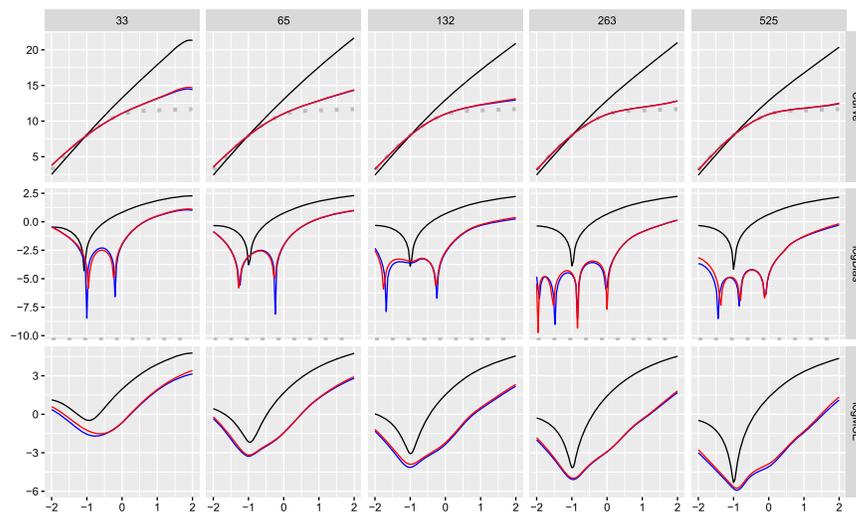


FIG 1. The marginal estimate of $\mu = f(z_1)$ for 'spouse' P2 using the full sample under scenario (S1). Compares the (true) population curve (broken gray) to the whole sample with equal weights (black), final analysis or 'marginal' weights (blue/dark gray), and household pairwise or 'second order' weights (red/light gray). Top to bottom: estimated curve, log of absolute bias, log of mean square error. Left to right: doubling of sample size for whole sample (100 to 1600).

adjust for informative sub-sampling of joint spouse-spouse pairs to more fully remove bias than marginal weighting. By contrast, the marginal weight for each individual is constructed based on quantities observed in the entire population, so it does not change or adapt to the particular sub-sample needed to study a conditioning event not observed in the population (other spouse's self-reported behavior); that is, the marginal weight for each individual is fixed to the same value for every sample or pair sub-sample that includes this individual (independent of the inclusion of the other spouse in the sample). While the marginal weighting scheme does demonstrate a notable improvement in estimation bias and MSE compared to the unweighted case, much of this improvement may be due to the informativeness of the first two stages of sampling. Using the marginal weights may lead to different, and potentially incorrect, inferential conclusions about the sub-population, as we will demonstrate in the Application section 6 that follows.

6. Application to NSDUH

Figure 3 shows the estimated relationship between the median ($q = 0.5$) number of days of alcohol use in the past month vs. age for the U.S. in 2014 from scenario (S2), which are the conditional curves based on the self-reported behavior of the spouse (in their sampled response). The top set of curves corresponds to the median estimates for individuals ($n \approx 4,000$) given that their spouse

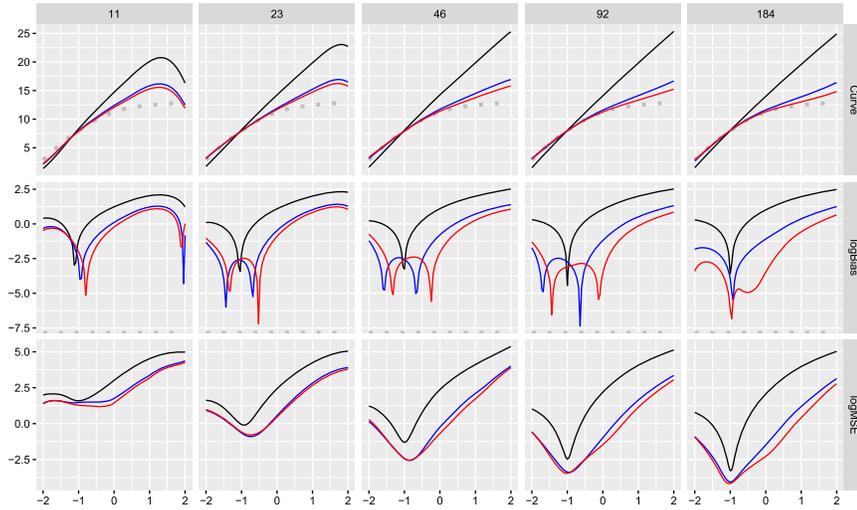


FIG 2. The conditional estimate of $\mu = f(z_1)$ for ‘spouse’ P2 given observed $x \geq 10$ for ‘other spouse’ P1 under scenario (S2). Compares the (true) population curve (broken gray) to the P1-P2 pair sample with equal weights (black), final analysis or ‘marginal’ weights (blue/dark gray), and household pairwise or ‘second order’ weights (red/light gray). Top to bottom: estimated curve, log of absolute bias, log of mean square error. Left to right: doubling of sample size for whole sample (100 to 1600).

reported any past month alcohol use ($x \geq 1$). The bottom set of curves represent the complement corresponding to median estimates for individuals ($n \approx 3,000$) whose spouse reported no past month alcohol use ($x = 0$). While the bottom curves show little difference between the weighting methods and little change across age, the top set of curves for the household pairwise weights suggest a different pattern than the marginal and equal weights. While the latter (marginal and equal weights) show a general lack of change after an initial increase in younger ages, the former (household pairwise weights) suggest a continued increase across age, perhaps with a jump around middle age; that is, alcohol consumption continues to increase with age for individuals who spouses consume alcohol at least once per month (It is important to remember that this study is cross-sectional and differences between ages may be due to cohort effects rather than a progression over time).

The agreement of the marginal weighted curves with those under equal weighting may suggest that the sampling of individuals is only weakly informative for marginal models of behavior, if we flexibly account for age in our models (Recall that sample is allocated disproportionately across age groups and states). However, the joint sampling of spouse-spouse pairs based on their pairs of ages is still informative with respect to joint or conditional models for behavior. Thus the use of the household pair weights removes bias associated with the informative joint sampling of different pairs of ages within each household. The marginal and equal weights over-emphasize respondents with younger spouses.

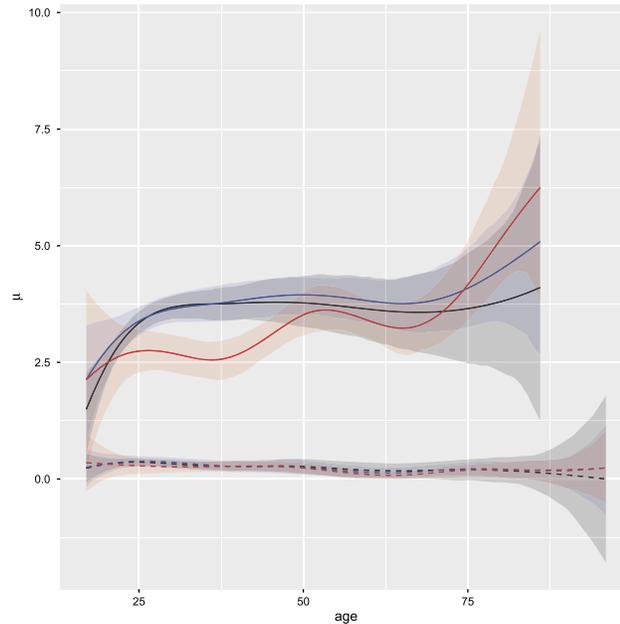


FIG 3. Estimated median and 95% intervals for number of days using alcohol in the past month among individuals conditional on their spouse's past month use of alcohol (based on the observed pair sample) under scenario (S2) using equal weights (black), final analysis or 'marginal' weights (blue/dark gray), and household pairwise or 'second order' weights (red/light gray). Solid lines indicate past month use by spouse; broken lines indicate no past month use by spouse.

These pairs are selected more often into the sample and they are more likely to have spouses who drink more frequently.

6.1. Application to other surveys

The use of household pairwise weights is most applicable to multi-stage cluster sample designs, particularly when the final cluster is small (e.g. household) and when individuals may have related behaviors, such as substance use (NSDUH) or consumer purchase patterns (CE). Then the potential for joint behaviors combined with joint differential selection could result in bias for conditional sub-population estimates which is not fully corrected by marginal weights. Pairwise weights are generally not useful for single stage designs, such as the Job Openings and Labor Turnover survey (JOLTS) example in Savitsky & Toth (2016). In this case, the joint probabilities of selection attenuate between units. Although these probabilities could be calculated in principle, for even moderate sample sizes this would be intractable due to the combinatorics required. Furthermore, many joint inclusions may be so small that the inverse weights will be large and unstable. Likewise, a multi-stage cluster design with large final cluster popu-

lations (e.g. cities) would have pairwise inclusions probabilities which rapidly attenuate similar to one-stage designs. More importantly, the joint measures available in these designs are typically not analytically important, essentially only the relationships of complete strangers are captured.

Final pairs weights used for analysis are best produced by the agency conducting the survey. These weights undergo non-response and post-stratification adjustments and have pair relationship p specific ‘multiplicity’ adjustments $N_{p_j} - 1$ within each household j . For example, the pair weights for a spouse-spouse pair in household j is normalized using $N_{p_j} = 2$, but the pair weights for a sibling-sibling pair may use $N_{p_j} = 3$ if there are three siblings. The analyst then selects the relationship specific weight (e.g. spouse-spouse) to perform inference. The NSDUH currently produces these adjustments for about a dozen household relationships (Westlake et al., 2016). We hope to influence other agencies to use similar methods to produce relationship specific pair weights which are a product of the base pair weight and the multiplicity adjustment. Access to these weights is typically restricted due to confidentiality concerns but may be facilitated through a research data center or similar mechanism.

7. Conclusions

This paper extends the previous work of Savitsky & Toth (2016) to include sampling designs in which the second order dependency between units does not fully attenuate. For multi-stage surveys, such as the CE and the NSDUH, the dependence structure between most units is dominated by the nearly independent first stages of selection. The dependency within final stage clusters is often negligible when estimating marginal models. Thus the marginal weights are typically robust for inference on the general population.

It is when we (i) sub-select within the last stage of selection (e.g., household) to target inference to an associated sub-population, (ii) include only joint or conditional sample responses between members within the last stage cluster for modeling, and (iii) the sub-selection probabilities are informative (related to the outcome (e.g. size or age)) that we achieve a gain from using the second order weights. Targeting inference to sub-populations of individuals in a household based on the behaviors of other members within the household, for example, may be of inferential interest to policy makers and researchers, but unbiased estimations for such sub-populations are not possible with first order weights.

When the inference targets the general population of individuals, however, while both first and second order weighting methods produce asymptotically unbiased inference about the population from estimation on realized samples, there may be a minor loss of efficiency for second order relative to first order weighting if the variance of the second order weights is larger than that for the first order weights. The more practical concern is the often lack of availability of second order weights. The burden to compute and store last stage (household) pairwise weights is much reduced because it is confined to the last stage of selection, so that the weight for each individual may be constructed from a single

(or, more generally, a few) pairwise term. To the extent that stakeholders express interest to conduct inference on sub-populations of individuals conditioned on the behavior of other individuals within the last stage, this paper may help encourage statistical agencies to pursue methods to provide access to second order weights while addressing potential concerns of confidentiality.

Appendix A: Enabling lemmas

Lemma A.1. *Suppose conditions (A1) and (A4) hold. Then for every $\xi > \xi_{N_\nu}$, a constant, $K > 0$, and any constant, $\delta > 0$,*

$$\mathbb{E}_{P_0, P_\nu} \left[\int_{P \in \mathcal{P} \setminus \mathcal{P}_{N_\nu}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \right] \leq \Pi(\mathcal{P} \setminus \mathcal{P}_{N_\nu}) \quad (17)$$

$$\mathbb{E}_{P_0, P_\nu} \left[\int_{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}^\pi(P, P_0) > \delta \xi} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \right] \leq 2\gamma \exp\left(\frac{-Kn_\nu \delta^2 \xi^2}{\gamma}\right). \quad (18)$$

The constant multiplier, $\gamma \geq 1$, defined in condition (A4), restricts the distribution of the sampling design by bounding all marginal inclusion probabilities for population units away from 0. As with the main result, the upper bound is increased by γ .

Proof. We begin by achieving the intermediate bound of Equation 32 in Savitsky & Toth (2016), on any set $B \in \mathcal{P}$, by replacing $1/\pi_{\nu\ell}$ in Equations 29 – 31 of Savitsky & Toth (2016) with $(1/(n_\nu - 1)) \sum_{k \neq \ell \in \delta_\nu} 1/\pi_{\nu\ell k}$, where δ_ν represents a particular sample of units, of fixed size n_ν , drawn from the space of samples, Δ_ν . The upper bound result in Equation 32 is achieved because each term, $1/\pi_{\nu\ell k}$, in the sum of $n_\nu - 1$ terms is greater than or equal to 1, such that the term, as a whole, is greater than or equal to 1 for each unit, $\ell \in \delta_\nu^*$, where δ_ν^* is that sample realization that maximizes $p/p_0(\mathbf{X}_\ell)$ (which is less than or equal to 1, by construction, however this can also be relaxed (Savitsky & Srivastava, 2016)). The bound specified in Equation 17 then directly follows from application of the intermediate bound on the set $\mathcal{P} \setminus \mathcal{P}_{N_\nu}$.

We next achieve an upper bound result for the expectation in Equation 18 stated in Equations 36 of Savitsky & Toth (2016) on models, P , in the slice, $\mathcal{A}_r^\pi = \{P \in \mathcal{P}_{N_\nu} : r\epsilon_{N_\nu} \leq d_{N_\nu}^\pi(P, P_0) \leq 2r\epsilon_{N_\nu}\}$ for integers, r , by again using the intermediate bound of Equation 32 in Savitsky & Toth (2016). This result on a slice derives from establishing an upper bound the (pairwise) sampling weighted, pseudo Hellinger distance, as follows:

$$d_{N_\nu}^{\pi, 2}(p_1, p_2) = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{1}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i))$$

$$\begin{aligned} &\leq \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \left[\frac{1}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{1}{\pi_{\nu ik}} \right] d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i)) \\ &\leq \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \left[\frac{1}{(N_\nu - 1)} (N_\nu - 1) \gamma d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i)) \right] \\ &\leq \gamma d_{N_\nu}^2(P_1, P_2), \end{aligned}$$

which, in turn, produces the upper bound stated in Equation 35 of Savitsky & Toth (2016) that directly leads to the result in Equation 36. Finally, the result of Equation 18 is directly achieved by adding up this upper bound on a slice over the countable collection of (dyadic) slices that form the set over which the integral is taken in Equation 18, as is outlined in the remainder of the proof in Savitsky & Toth (2016). This concludes the proof. \square

Lemma A.2. For every $\xi > 0$ and measure Π on the set,

$$B = \left\{ P : -P_0 \log \left(\frac{p}{p_0} \right) \leq \xi^2, P_0 \left(\log \frac{p}{p_0} \right)^2 \leq \xi^2 \right\}$$

under the conditions (A2), (A3), (A4), (A5), (A6), we have for every $C > 0$, $C_3 = C_4 + C_5 + 1$ and N_ν sufficiently large,

$$Pr \left\{ \int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi}(\mathbf{X}_i \delta_{\nu i}) d\Pi(P) \leq \exp[-(1 + C)N_\nu \xi^2] \right\} \leq \frac{\gamma + C_3}{C^2 N_\nu \xi^2}, \quad (19)$$

where the above probability is taken with the respect to P_0 and the sampling generating distribution, P_ν , jointly.

The bound of “1” in the numerator of the result for Lemma 8.1 of Ghosal et al. (2000), is replaced with $\gamma + C_3$ for our generalization of this result in Equation 19. The sum of positive constants, $\gamma + C_3$, is greater than 1 and will be larger for sampling designs where the pairwise inclusion probabilities, $\{\pi_{\nu ij}\}$, express a relatively larger variation, which will tend to produce samples that are less representative of the underlying population.

Proof. The proof exactly follows that of Savitsky & Toth (2016) by bounding the probability expression on left-hand size of Equation 19 with,

$$\begin{aligned} &Pr \left\{ \mathbb{G}_{N_\nu}^\pi \int_{P \in \mathcal{P}} \log \frac{p}{p_0} d\Pi(P) \leq -\sqrt{N_\nu} \xi^2 C \right\} \\ &\leq \frac{\int_{P \in \mathcal{P}} \left[\mathbb{E}_{P_0, P_\nu} \left(\mathbb{G}_{N_\nu}^\pi \log \frac{p}{p_0} \right)^2 \right] d\Pi(P)}{N_\nu \xi^4 C^2}, \end{aligned} \quad (20)$$

where we have used Chebyshev to achieve the right-hand bound of Equation 20. We now proceed to further bound the numerator in the right-hand side of Equation 20, which will result in the expression on the right-hand side of Equation 19. The expectation inside the square brackets on the right-hand side of Equation 20 is taken with respect to the joint distribution of population generation and the taking of a sample. In the sequel, define $\mathcal{A}_\nu = \sigma(\mathbf{X}_1, \dots, \mathbf{X}_{N_\nu})$ as the sigma field of information potentially available for the N_ν units in population, U_ν .

$$\mathbb{E}_{P_0, P_\nu} \left[\mathbb{G}_{N_\nu}^\pi \log \frac{p}{p_0} \right]^2 \tag{21a}$$

$$= \mathbb{E}_{P_0, P_\nu} \left[\sqrt{N_\nu} (\mathbb{P}_{N_\nu}^\pi - \mathbb{P}_{N_\nu}) \log \frac{p}{p_0} - \sqrt{N_\nu} (\mathbb{P}_0 - \mathbb{P}_{N_\nu}) \log \frac{p}{p_0} \right]^2 \tag{21b}$$

$$= \mathbb{E}_{P_0, P_\nu} \left[\sqrt{N_\nu} (\mathbb{P}_{N_\nu}^\pi - \mathbb{P}_{N_\nu}) \log \frac{p}{p_0} - \sqrt{N_\nu} \mathbb{G}_{N_\nu} \log \frac{p}{p_0} \right]^2 \tag{21c}$$

$$\leq N_\nu \mathbb{E}_{P_0, P_\nu} \left[(\mathbb{P}_{N_\nu}^\pi - \mathbb{P}_{N_\nu}) \log \frac{p}{p_0} \right]^2 + \mathbb{E}_{P_0} \left[\mathbb{G}_{N_\nu} \log \frac{p}{p_0} \right]^2 \tag{21d}$$

$$\leq N_\nu \mathbb{E}_{P_0, P_\nu} \left[(\mathbb{P}_{N_\nu}^\pi - \mathbb{P}_{N_\nu}) \log \frac{p}{p_0} \right]^2 + \xi^2, \tag{21e}$$

where the bound of the expectation of the centered empirical process approximation over the units in the population taken with respect to the population generating distribution, included in the second term in Equation 21d, is shown to be bounded from above (for any constant, $C > 0$) under Lemma B.2 of Savitsky & Toth (2016) by replacing $(\gamma + C_3)$ with “1” in the bound $\xi^2(\gamma + C_3)$ because we draw a finite population from P_0 and do not take a further informative sample under P_ν .

We now proceed to further simplify the bound in the first term of Equation 21d.

$$N_\nu \mathbb{E}_{P_0, P_\nu} \left[(\mathbb{P}_{N_\nu}^\pi - \mathbb{P}_{N_\nu}) \log \frac{p}{p_0} \right]^2 \tag{22a}$$

$$= N_\nu \mathbb{E}_{P_0, P_\nu} \left[\frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \left(\left\{ \frac{1}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} \right\} - 1 \right) \log \frac{p}{p_0} (\mathbf{X}_i) \right]^2 \tag{22b}$$

$$= \frac{1}{N_\nu} \sum_{i, j \in U_\nu} \mathbb{E}_{P_0, P_\nu} \left[\left(\left\{ \frac{1}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} \right\} - 1 \right) \times \left(\left\{ \frac{1}{(N_\nu - 1)} \sum_{\ell \neq j \in U_\nu} \frac{\delta_{\nu j} \delta_{\nu \ell}}{\pi_{\nu j \ell}} \right\} - 1 \right) \log \frac{p}{p_0} (\mathbf{X}_i) \frac{p}{p_0} (\mathbf{X}_j) \right]^2 \tag{22c}$$

$$= \frac{1}{N_\nu} \sum_{i \neq j \in U_\nu} \mathbb{E}_{P_0} \left[\mathbb{E}_{P_\nu} \left\{ \frac{1}{(N_\nu - 1)^2} \sum_{k \neq i, \ell \neq j \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k} \delta_{\nu j} \delta_{\nu \ell}}{\pi_{\nu ik} \pi_{\nu j \ell}} \right. \right.$$

$$\begin{aligned}
 & + \frac{1}{(N_\nu - 1)^2} \sum_{k \neq i, k \neq j \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu j} \delta_{\nu k}}{\pi_{\nu ik} \pi_{\nu jk}} \\
 & - \frac{1}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} \\
 & - \frac{1}{(N_\nu - 1)} \sum_{\ell \neq j \in U_\nu} \left. \frac{\delta_{\nu j} \delta_{\nu \ell}}{\pi_{\nu j \ell}} + 1 \right| \mathcal{A}_\nu \left. \left(\log \frac{p}{p_0}(\mathbf{X}_i) \frac{p}{p_0}(\mathbf{X}_j) \right) \right] \\
 & + \frac{1}{N_\nu} \sum_{i=j \in U_\nu} \mathbb{E}_{P_0} \left[\mathbb{E}_{P_\nu} \left\{ \frac{1}{(N_\nu - 1)^2} \sum_{k \neq \ell \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k} \delta_{\nu \ell}}{\pi_{\nu ik} \pi_{\nu i \ell}} \right. \right. \\
 & \quad \left. \left. - \frac{2}{(N_\nu - 1)} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}} + 1 \right| \mathcal{A}_\nu \right. \left. \left(\log \frac{p}{p_0}(\mathbf{X}_i) \right)^2 \right] \\
 & + \frac{1}{N_\nu} \sum_{i=j \in U_\nu} \mathbb{E}_{P_0} \left[\mathbb{E}_{P_\nu} \left\{ \frac{1}{(N_\nu - 1)^2} \sum_{k \neq i \in U_\nu} \frac{\delta_{\nu i} \delta_{\nu k}}{\pi_{\nu ik}^2} \right| \mathcal{A}_\nu \right. \left. \left(\log \frac{p}{p_0}(\mathbf{X}_i) \right)^2 \right] \tag{22d}
 \end{aligned}$$

$$\begin{aligned}
 & = \frac{1}{N_\nu} \sum_{i \neq j \in U_\nu} \mathbb{E}_{P_0} \left[\left\{ \frac{1}{(N_\nu - 1)^2} \sum_{k \neq i, \ell \neq j \in U_\nu} \frac{\pi_{\nu i k j \ell}}{\pi_{\nu ik} \pi_{\nu j \ell}} \right. \right. \\
 & \quad \left. \left. - 1 + \frac{1}{(N_\nu - 1)^2} \sum_{k \neq i, k \neq j \in U_\nu} \frac{\pi_{\nu i j k}}{\pi_{\nu ik} \pi_{\nu j k}} \right\} \left(\log \frac{p}{p_0}(\mathbf{X}_i) \frac{p}{p_0}(\mathbf{X}_j) \right) \right] \\
 & + \frac{1}{N_\nu} \sum_{i=j \in U_\nu} \mathbb{E}_{P_0} \left[\left\{ \frac{1}{(N_\nu - 1)^2} \sum_{k \neq \ell \neq i \in U_\nu} \frac{\pi_{\nu i k \ell}}{\pi_{\nu ik} \pi_{\nu i \ell}} - 1 \right\} \left(\log \frac{p}{p_0}(\mathbf{X}_i) \right)^2 \right] \\
 & + \frac{1}{N_\nu} \sum_{i=j \in U_\nu} \mathbb{E}_{P_0} \left[\left\{ \frac{1}{(N_\nu - 1)^2} \sum_{k \neq i \in U_\nu} \frac{1}{\pi_{\nu ik}} \right\} \left(\log \frac{p}{p_0}(\mathbf{X}_i) \right)^2 \right] \tag{22e}
 \end{aligned}$$

$$\begin{aligned}
 & \leq (N_\nu - 1) \sup_\nu \max_{i, j, k, \ell: i \neq j, k \neq i, \ell \neq j \in U_\nu} \left| \frac{\pi_{\nu i k j \ell}}{\pi_{\nu ik} \pi_{\nu j \ell}} - 1 \right| \left\{ \mathbb{E}_{P_0} \log \frac{p}{p_0}(\mathbf{X}_i) \frac{p}{p_0}(\mathbf{X}_j) \right\} \\
 & + \sup_\nu \max_{i, k, \ell: k \neq \ell \neq i \in U_\nu} \left| \frac{\pi_{\nu k \ell i}}{\pi_{\nu ki} \pi_{\nu \ell i}} \right| \left\{ \mathbb{E}_{P_0} \log \frac{p}{p_0}(\mathbf{X}_i)^2 \right\} \\
 & + \sup_\nu \left[\frac{1}{(N_\nu - 1)} \frac{1}{\min_{i, k: k \neq i \in U_\nu} |\pi_{\nu ik}|} \right] \left\{ \mathbb{E}_{P_0} \log \frac{p}{p_0}(\mathbf{X}_i)^2 \right\} \leq (C_4 + C_5 + \gamma) \xi^2, \tag{22f}
 \end{aligned}$$

for sufficiently large N_ν , where we have applied the condition for $P \in B$ in each of the three terms in the last inequality and conditions (A6), (A5) and (A4) for each term in the last inequality, from left-to-right. We additionally note that

$\pi_{\nu ik\ell} = \pi_{\nu ik}$ when $\ell = k$, $\ell, k \in U_\nu$ and denote $\pi_{\nu k|i} := \mathbb{E}_{P_\nu}(\delta_{\nu k} | \delta_{\nu i} = 1)$.

We may finally bound the expectation on the right-hand side of Equation 20,

$$\begin{aligned} \mathbb{E}_{P_0, P_\nu} \left[\mathbb{G}_{N_\nu}^\pi \log \frac{p}{p_0} \right]^2 &\leq (C_4 + C_5 + \gamma)\xi^2 \\ &+ \xi^2 \leq (C_4 + C_5 + \gamma + 1)\xi^2 \leq (C_3 + \gamma)\xi^2, \end{aligned} \quad (23)$$

for N_ν sufficiently large, where we set $C_3 := C_4 + C_5 + 1$. This concludes the proof. \square

Appendix B: Simulation details

This appendix contains detailed information on the population model (Section B.1) and sampling design (Section B.2) used to generate the simulation results for the main text. It also describes alternative methods of calculating second order or ‘pairwise’ sampling weights (Section B.3) and compares the distribution and properties of these weights (Section B.4) for the simulated population.

B.1. Example population model

The response x is drawn from an AL distribution with $q = 0.5$. We choose $\tau = 8$ to yield a relatively precise response. We let μ depend on two predictors z_1 and z_2 . The variable z_1 represents the observed information available for analysis, whereas z_2 represents information available for sampling, which is either ignored or not available for analysis. The z_1 and z_2 distributions for P1 and P3 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 $\tilde{z}_{2,ijk} = z_{2,ijk} - \min(z_{2,ijk}) + 1$ for $i = 1, \dots, 3$ individuals, $j = 1, \dots, 10$ HHs, and $k = 1, \dots, 200$ PSUs. The conditional quantile for P1 and P3 ($i = 1, 3$) within each HH j and PSU k is

$$\mu_i = 10 + 1z_{1,i} + 0.5z_{2,i} + 0.5z_{1,i}z_{2,i} - z_{1,i}^2$$

P2 is given a distribution for $z_{2,2jk}$ that depends on P1’s value for $z_{2,1jk}$ within HH j . Within each PSU, k , and HH, j , the distributions for z_2 ($i = 2$) are $z_{2,2}|z_{2,1} \sim \mathcal{E}(1/z_{2,1})$, so $E(z_{2,2}|z_{2,1}) = z_{2,1}$. The distribution for μ_2 is further set to depend on whether the z_2 value for P1 in the same HH is higher or lower than the median $Q_{0.5}$ of z_2 among the population of P1s.

$$\mu_2 = \begin{cases} 10 + 1z_{1,2} + 0.25z_{2,2} + 0.25z_{1,2}z_{2,2} - 2z_{1,2}^2; & z_{2,1} \leq Q_{0.5}(z_{2,1jk}) \\ 10 + 1z_{1,2} + 0.75z_{2,2} + 0.75z_{1,2}z_{2,2}; & z_{2,1} > Q_{0.5}(z_{2,1jk}) \end{cases}$$

In terms of P2-P1 and P2-P3 pairs within each HH, there are now different distributions for both the outcome $x_{2,jk}$ (via conditional μ_2) and the joint selection probability (via conditional size $z_{2,2jk}|z_{2,1jk}$) even though the marginal distributions for outcomes x_{1jk} , x_{3jk} and size measures $z_{2,1jk}$, $z_{2,3jk}$ are the same.

B.2. Sampling from the population

For the simulation, we choose a population size $N = 6000$, with 200 primary sampling units (PSUs) each containing 10 households (HHs). The number of selected PSUs was varied $K \in \{10, 20, 40, 80, 160\}$, the number of HHs within each PSU was fixed at 5, and the number of selected individuals within each HH was 2 (a pair). 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 $Z_{2,k} = \sum_{ij} z_{2,ij|k}$ 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 probability proportional to size (PPS) algorithm (Brewer, 1975).
2. Once PSUs are selected, for each HH within the selected PSUs indexed by j an aggregate size measure $Z_{2,j|k} = \sum_i z_{2,i|jk}$ 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 proportional to size based on Brewer's PPS algorithm.
3. Within each selected HH, a pair of persons (2 out of P1, P2, P3) is selected jointly. Firstly, all $\binom{3}{2} = 6$ pairs are given a size equal to the sum of the individual size measures. So $Z_{2,ii'|jk} = z_{2,i|jk} + z_{2,i'|jk}$. Secondly, a single pair is then directly selected with probability proportional to this size measure. Individual (marginal) probabilities of selection for each of P1, P2, and P3 can be computed directly from the 6 pair inclusion probabilities.

B.3. Calculation of second order weights

The calculation of second order weights can be motivated in different ways:

1. "Full" second order weights can be constructed by populating the full $\binom{n}{2}$ matrix of second order inclusion probabilities across the entire sample $\pi_{i,i'}$, taking their inverse $w_{i,i'}^{(2)} = 1/\pi_{i,i'}$, and then summing for each individual record and normalizing by the number of pairs in the target population $w_i^{(2f)} = \sum_{i'} w_{i,i'}^{(2)}/(N-1)$ where N is the population size. Because there are three stages, the joint weight component $w_{i,i'}^{(2)}$ is calculated differently depending on whether individuals i and i' are in the same HH, different HHs but same PSU, or different PSUs:
 - Case 1: i and i' are in different PSUs. Since we treat selection of PSUs as independent and we select HHs and persons separately across PSUs, the second order weight component is simply the product of the first order weights: $w_{i,i'}^{(2)} = w_i^{(1)} w_{i'}^{(1)}$.
 - Case 2: i and i' are in the same PSU but different HHs: Since we treat HHs as being conditionally independent given the selection of the PSUs, the second order weight term has a common PSU term and distinct HH and individual terms: $w_{i,i'}^{(2)} = w_1^k w_2^{j|k} w_3^{i|jk} w_2^{j'|k} w_3^{i'|j'k}$

- Case 3: i and i' are in the same HH and thus selected as a pair: The second order weight term has a common PSU term and common HH term and common person-pair term: $w_{i,i'}^{(2)} = w_1^k w_2^{j|k} w_3^{i,i'|jk}$
2. 'Last stage' or 'pair' weights can be constructed by assuming the HH is the unit of analysis. This means that first order weight components provide a consistent estimate at the HH level and any sampling dependence across households is negligible. Because responses are not available for all persons within a household, we then incorporate the joint sampling dependence of each pair selected. Instead of summing over all i, i' pairs in the sample, we only sum over the i, i' pairs in each HH, which in this example is a single pair. We use the household size (N_{p_j}) to normalize by the number of pairs within each household because each roster of the HH is treated as a population: $w_i^{(2p)} = w_{i,i'}^{(2)}/(N_{p_j} - 1) = w_1^k w_2^{j|k} w_3^{i,i'|jk}/(N_{p_j} - 1)$. This set of weights is appealing for several reasons. It is most likely that dependence in the response x is strongest within a HH rather than between HHs and PSUs. It is also possible to measure each N_p during data collection. It usually needs to be known before sampling can occur. Whereas the general N for a particular domain may not be available and must be estimated by summing the first order weights.
 3. 'Stagewise' second order weights can be constructed by assuming that each of the three stages is a conditionally independent sampling with conditional population frames of PSUs, HHs, and individuals. So any second order dependence would only need to be captured within each stage. Then the first order weights for each stage would be replaced by the scaled sum of the second order weights for that stage. For example the stage 1 weights for PSUs would be $w_1^{k,k'} = \sum_{k'} w_1^{k,k'}/(N_S - 1)$ with $w_1^{k,k'} = 1/\pi_{k,k'}$ and N_S is the number of PSUs in the population. Similarly $w_2^{j|k} = \sum_{j'} w_2^{j,j'|k}/(N_{h_k} - 1)$ with $w_2^{j,j'|k} = 1/\pi_{j,j'|k}$ and N_{h_k} is the number of HHs in PSU k . The within household stage weights would be $w_3^{i,i'|jk} = w_3^{i,i'|jk}/(N_{p_j} - 1)$. Then the final stagewise second order weight would be

$$w_i^{(2s)} = w_1^k w_2^{j|k} w_3^{i,i'|jk} = \left(\sum_{k'} w_1^{k,k'}/(N_S - 1) \right) \times \left(\sum_{j'} w_2^{j,j'|k}/(N_{h_k} - 1) \right) \times \left(w_3^{i,i'|jk}/(N_{p_j} - 1) \right)$$

For moderate to large samples, particularly for multi-stage designs with low dependence at the first levels of sampling, the four weights presented here ($w_i^{(1)}$, $w_i^{(2f)}$, $w_i^{(2p)}$, $w_i^{(2s)}$) effectively become only two distinct sets of weights ($w_i^{(1)}$ and $w_i^{(2p)}$). For full second order weights $w_i^{(2f)}$, the sum is dominated by Case 1. Then the sum $w_i^{(2f)} = \sum_{i'} w_{i,i'}^{(2)}/(N - 1) \rightarrow w_i^{(1)}(\sum_{i'} w_{i'}^{(1)})/(N - 1)$. The sum of the first order weights ($\sum_i w_i = \sum_{i'} w_{i'} + w_i$) converges to N so $w_i^{(2f)} \approx$

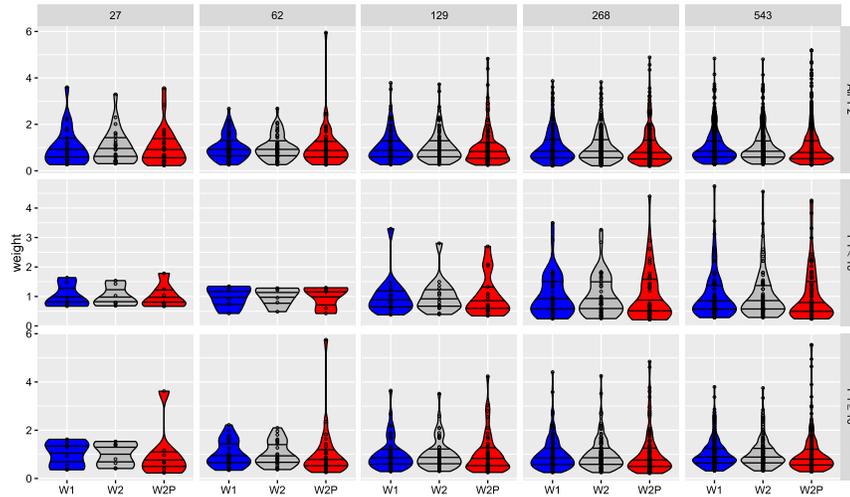


FIG 4. Distributions of marginal ($W1$), full second order ($W2$), and within HH second order ($W2P$) weights for $P2$ across realizations of different sample sizes (column heading) and by sub-domain (top to bottom) all $P2$, $P2$ where ‘spouse’ $P1$ ’s response < 10 , $P2$ where ‘spouse’ $P1$ ’s response ≥ 10

$\left(\frac{N-w_i^{(1)}}{N-1}\right) w_i^{(1)} \approx w_i^{(1)}$. For stagewise weights $w_i^{(2s)}$, the first and second stage components from the PPS design are typically assumed to be the products of independent samples. By a similar argument, for moderate sample sizes w_1^k , and $w_2^{j|k}$ is effectively the same as w_1^k and $w_2^{j|k}$. The third or last stage weights are still distinct, so the stagewise weights $w_i^{(2s)}$ are very similar to the last stage pair weights $w_i^{(2p)}$.

B.4. Example properties of pairwise weights

Figure 4 compares the distribution of sampling weights under marginal, full pairwise and within-household pairwise weighting, from left-to-right, within each plot panel. The panels in each column present distributions for realized samples of increasing size, from left-to-right. The rows compare the weight distributions for all $P2$ units, $P2$ units where spouse $P1$ ’s response < 10 , and $P2$ units where spouse $P1$ ’s response ≥ 10 , from top-to-bottom.

We observe that while the weight distributions are highly similar for marginal and full pairwise weighting, on the one hand, there are notable differences between the first two and household pairwise weighting, on the other hand. Figure 5 plots the distributions for the ratio of full and within household pairwise weight, to better understand the differences in their underlying distributions. Taken together, both figures reveal that only the household pairwise weights actually redistribute the marginal weight, whereas the full sample (second order) pairwise weights quickly collapse to the marginal weights. The full pairwise

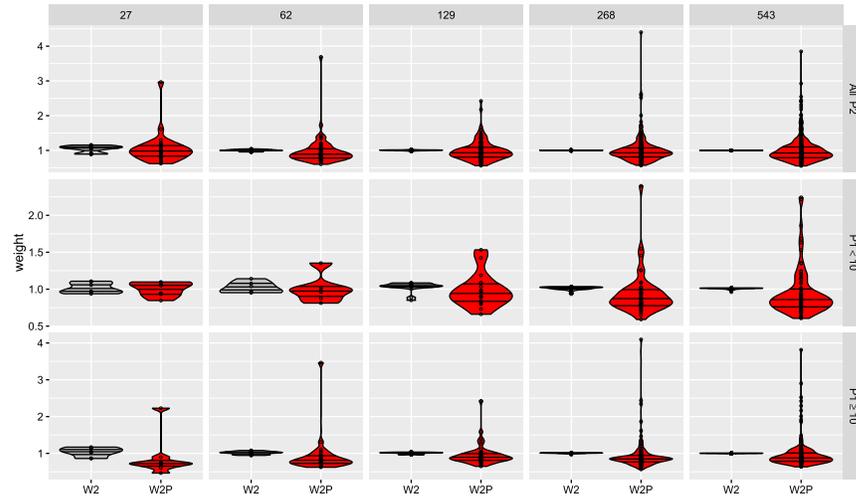


FIG 5. Distributions of the ratio of full second order ($W2$) and within HH second order ($W2P$) weights to marginal weights for $P2$ across realizations of different sample sizes (column heading) and by sub-domain (top to bottom) all $P2$, $P2$ where 'spouse' $P1$'s response < 10 , $P2$ where 'spouse' $P1$'s response ≥ 10

weights converge to the marginal weights because the majority of terms in each summation to construct a weight value for each individual are from pairings across different PSUs and HHs. These terms are dominated by the early, nearly independent sampling stages and thus the small number (only one for pair samples) of within HH components provide negligible contributions to the sum.

References

- Breslow, N. E. & 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](#)
- Brewer, K. (1975), 'A simple procedure for π pswor', *Australian Journal of Statistics* **17**, 166–172. [MR0440754](#)
- Carpenter, B. (2015), 'Stan: A probabilistic programming language', *Journal of Statistical Software* **76**(1).
- Clifford, S. & Choy, S. L. (2012), Bayesian splines, in C. L. Alston, K. L. Mengersen & A. N. Pettitt, eds, 'Case Studies in Bayesian Statistical Modelling and Analysis', Wiley series in probability and statistics, John Wiley & Sons, Ltd, Oxford, pp. 197–220.
- Dong, Q., Elliott, M. R. & Raghunathan, T. E. (2014), 'A nonparametric method to generate synthetic populations to adjust for complex sampling design features', *Survey Methodology* **40**(1), 29–46.
- Ghosal, S., Ghosh, J. K. & Vaart, A. W. V. D. (2000), 'Convergence rates of posterior distributions', *Ann. Statist* **28**(2), 500–531. [MR1790007](#)

- Ghosal, S. & van der Vaart, A. (2007), ‘Convergence rates of posterior distributions for noniid observations’, *Ann. Statist.* **35**(1), 192–223. [MR2332274](#)
- Kunihama, T., Herring, A. H., Halpern, C. T. & Dunson, D. B. (2014), Non-parametric bayes modeling with sample survey weights, *Statist. Probab. Lett* **113**, 41–48. [MR3480393](#)
- Morton, K. B., Aldworth, J., Hirsch, E. L., Martin, P. C. & Shook-Sa, B. E. (2016), Section 2, sample design report, in ‘2014 National Survey on Drug Use and Health: Methodological Resource Book’, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.
- Rao, J. N. K. & Wu, C. (2010), ‘Bayesian pseudo-empirical-likelihood intervals for complex surveys’, *Journal of the Royal Statistical Society Series B* **72**(4), 533–544. [MR2758527](#)
- Reed, C. & Yu, K. (2009), A partially collapsed gibbs sampler for bayesian quantile regression, Technical report. [MR2749833](#)
- Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC. [MR2130347](#)
- Savitsky, T. D. & Srivastava, S. (2018), ‘Scalable Bayes under Informative Sampling’, *Scandinavian Journal of Statistics*.
URL: <https://doi.org/10.1111/sjos.12312>
- Savitsky, T. D. & Toth, D. (2016), ‘Bayesian Estimation Under Informative Sampling’, *Electronic Journal of Statistics* **10**(1), 1677–1708. [MR3522657](#)
- Si, Y., Pillai, N. S. & Gelman, A. (2015), ‘Bayesian nonparametric weighted sampling inference’, *Bayesian Anal.* **10**(3), 605–625. [MR3420817](#)
- Westlake, M., Chen, P., Gordek, H., Williams, M. & Hughes, A. (2016), Section 12, questionnaire dwelling unit-level and person pair-level sampling weight calibration, in ‘2014 National Survey on Drug Use and Health: Methodological Resource Book’, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.
- Yi, G. Y., Rao, J. N. K. & Li, H. (2016), ‘A Weighted Composite Likelihood Approach for Analysis of Survey Data under Two-level Models’, *Statistica Sinica* **26**, 569–587. [MR3497760](#)