Published in final edited form as:

Stat Probab Lett. 2016 June ; 113: 41–48. doi:10.1016/j.spl.2016.02.009.

Nonparametric Bayes modeling with sample survey weights

T. Kuniyama^a, A. H. Herring^c, C. T. Halpern^d, D. B. Dunson^b

^aDepartment of Statistics, University of Washington, Seattle, WA 98195, U.S.A.

^bDepartment of Statistical Science, Duke University, Durham, NC 27708, U.S.A.

^cDepartment of Biostatistics and Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

^dDepartment of Maternal and Child Health and Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

Abstract

In population studies, it is standard to sample data via designs in which the population is divided into strata, with the different strata assigned different probabilities of inclusion. Although there have been some proposals for including sample survey weights into Bayesian analyses, existing methods require complex models or ignore the stratified design underlying the survey weights. We propose a simple approach based on modeling the distribution of the selected sample as a mixture, with the mixture weights appropriately adjusted, while accounting for uncertainty in the adjustment. We focus for simplicity on Dirichlet process mixtures but the proposed approach can be applied more broadly. We sketch a simple Markov chain Monte Carlo algorithm for computation, and assess the approach via simulations and an application.

Keywords

Biased sampling; Dirichlet process; Mixture model; Stratified sampling; Survey data

1. Introduction

In sample surveys, it is routine to conduct stratified sampling designs to ensure that a broad variety of groups are adequately represented in the sample. In particular, the population is divided into mutually exclusive strata having different probabilities of inclusion. Data from stratified probability designs cannot be analyzed as if they are a random sample from the super-population without potentially large amounts of bias. To correct for discrepancies in the statistical analysis, survey weights are constructed. There is a rich literature on including adjustments for stratified sampling designs in estimation. However, the vast majority of such

tsuyoshi.kuniyama@uw.edu (T. Kuniyama).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

methods are not appropriate in model-based inferences, particularly under nonparametric Bayes frameworks.

Little (2004) and Gelman (2007) clarify the importance of including survey weights into model-based analyses. Zheng and Little (2003, 2005) propose a nonparametric spline model and Chen et al. (2010) extend the framework for binary variables. Although these approaches can flexibly connect the survey weights with the response, they rely on the assumption of survey weights being known for all population units. Zangeneh and Little (2012) propose a modification to allow the number of non-sampled units to be unknown. Si et al. (2015) instead propose a nonparametric model in which the survey weights are linked with a response through a Gaussian process regression. However, additional modeling of survey weights for non-sampled subjects in the population can lead to highly complex models.

In this article, we propose a simple approach in which we apply standard mixture models, such as Dirichlet process mixtures, for the selected sample, and then adjust the mixture weights based on the survey weights. We allow probabilistic uncertainty in this adjustment in a Bayesian manner. Posterior computation relies on a simple modification to add an additional step to Markov chain Monte Carlo algorithms for mixture models.

2. Mixture Models with Survey Weights

2.1. Adjusted density estimates

Let $y_1, \dots, y_N$ denote independently and identically distributed observations from a superpopulation density f_0 with $y_i \in \mathcal{R}$ for $i \in D = \{1, \dots, N\}$. From the finite population D , n subjects are sampled, with $w_i = c/\pi_i$ the survey weight for subject i , c a positive constant, and π_i the inclusion probability for $i \in D$. We assume D can be divided into mutually exclusive subpopulations $D_1, \dots, D_M$, with $\{y_i, i \in D_m\}$ independently and identically distributed from density f_m for $m = 1, \dots, M$. Then, f_0 can be expressed as

$$f_0(y) = \sum_{m=1}^M \nu_m f_m(y), \quad (1)$$

where $\nu_m \geq 0$ and $\sum_{m=1}^M \nu_m = 1$. By applying kernel density estimation to each f_m in (1), Buskirk (1998) and Bellhouse and Stafford (1999) propose an adjusted density estimate,

$$\hat{f}_0(y) = \sum_{i \in S} \frac{\tilde{w}_i}{b} \mathcal{K}\left(\frac{y - y_i}{b}\right), \quad (2)$$

where $S \subset D$ are the selected subjects in the survey, $\tilde{w}_i = w_i / \sum_{j \in S} w_j$, $\mathcal{K}$ is a kernel function and $b > 0$. Estimator (2) adjusts for bias in the usual kernel estimator applied to sample S by

modifying the weight for the i th subject from $1/n$ to $\tilde{w}_i$. This adjustment leads to consistency under some conditions (Buskirk and Lohr (2005)).

2.2. Bayesian adjustments with uncertainty

Section 2.1 focuses on univariate continuous variables, while our goal is to develop a general approach for adjusting posterior distributions to take into account sample survey weights. Let $y \in \mathcal{Y}$ denote a random variable, with $\mathcal{Y}$ a Polish space that may correspond to a p -dimensional Euclidean space, a discrete space, a mixed continuous and discrete space, a non-Euclidean Riemannian manifold, such as a sphere, and other cases. Extending (1) to general spaces, we let $f_0(\cdot)$ and $f_m(\cdot)$, for $m = 1, \dots, M$, denote densities on $\mathcal{Y}$ with respect to a dominating measure μ . The density in the m th subpopulation is expressed as a mixture,

$$f_m(y) = \sum_{h=1}^H \nu_{mh} f(y|\theta_h), \quad (3)$$

where $\nu_{mh} \geq 0$, $\sum_{h=1}^H \nu_{mh} = 1$ and θ_h are parameters characterizing the h th mixture component. Then, f_0 can be approximately expressed as a mixture having the same kernels as in (3) but with adjusted weights as in (2).

Theorem 1. Let $S_i \in \{1, \dots, H\}$ denote the mixture index for subject i for $i \in S$. Let $S_h = \{i : s_i = h, i \in S\}$, for $h = 1, \dots, H$. Then, for large N and n ,

$$f_0(y) \approx \sum_{h=1}^H \frac{\sum_{i \in S_h} w_i^{1/c}}{N} f(y|\theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y|\theta_{s_i}). \quad (4)$$

PROOF. Letting N_m be the number of subjects in D_m , $N_m/N \rightarrow \nu_m$ as $N \rightarrow \infty$ by the law of large numbers. Letting w_m^* and π_m^* denote the survey weight and inclusion probability for the m th subpopulation, $w_i = w_m^*$ and $\pi_i = \pi_m^*$ for $i \in D_m$. From (1) and (3), f_0 can be expressed as

$$\begin{aligned} f_0(y) &= \sum_{m=1}^M \nu_m f_m(y) \approx \sum_{m=1}^M \frac{N_m}{N} f_m(y) = \sum_{h=1}^H \sum_{m=1}^M \frac{N_m \nu_{mh}}{N} f(y|\theta_h) \\ &\approx \sum_{h=1}^H \frac{\sum_{i \in S_h} w_i^{1/c}}{N} f(y|\theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y|\theta_{s_i}). \end{aligned} \quad (5)$$

The first approximation in (5) can be induced by $N_m \approx w_m^* n_m / c$ and

$$\nu_{mh} \approx \frac{\sum_{i \in S} 1(i \in D_{mh})}{n_m},$$

for large N_m and n_m , where $D_{mh} = \{i: s_i = h, i \in D_m\}$. The second approximation in (5) is based on $c \approx \sum_{i \in S} w_i / N$, which is derived by summation of $N_m \approx w_m^* n_m / c$ over m .

Under random designs with $w_i \propto c$, f_0 can be approximated by

$$f_R(y) = \sum_{h=1}^H \frac{\sum_{i \in D} 1(i \in S_h)}{n} f(y | \theta_h) = \sum_{i \in S} \frac{1}{n} f(y | \theta_{s_i}). \quad (6)$$

Comparing the last terms in (4) and (6), we can interpret that the bias can be adjusted by shifting the weight for the i th sampled subject from $1/n$ to $\tilde{w}_i$ as in (2).

We propose a simple Bayesian adjustment method using the result in Theorem 1. Although survey weights are treated as random variables in Bayesian settings, additional modeling of the weights can induce highly complex models. Instead, we avoid such complexity in the model but still incorporate uncertainty by generating mixture weights. We consider a standard Bayesian mixture model,

$$f_B(y) = \sum_{h=1}^H \lambda_h f(y | \theta_h), \quad \lambda \sim \pi(\lambda), \quad \theta_h \sim \pi(\theta_h), \quad (7)$$

where $\lambda = (\lambda_1, \dots, \lambda_H)'$ with $\lambda_h \geq 0$ and $\sum_{h=1}^H \lambda_h = 1$, and $\pi(\lambda)$ and $\pi(\theta_h)$ are priors for λ and θ_h . For example, using a truncated stick-breaking process (Ishwaran and James (2001)), we let $\lambda_h = V_h \prod_{k < h} (1 - V_k)$, $V_h \sim \text{Beta}(1, \alpha)$ for $h = 1, \dots, H-1$ with $V_H = 1$. However, our focus is not on the specific mixture model and prior but on the adjustment for sampling bias, and alternative priors can be used without complication.

Comparing the second terms in (4) and (6), the difference is in the mixture weights. The expression in (4) can be interpreted as implying that $\sum_{i \in S_h} w_i / c$ subjects are generated from the h th mixture component in the population. Updating the prior $\tilde{\lambda} \sim \text{Dirichlet}(a_1, \dots, a_H)$ with this information, we obtain the following conditional posterior distribution for the adjusted weights $\tilde{\lambda} = (\tilde{\lambda}_1, \dots, \tilde{\lambda}_H)'$,

$$\tilde{\lambda} \sim \text{Dirichlet}\left(a_1 + \frac{1}{\tilde{c}} \sum_{i: s_i = 1} w_i, \dots, a_H + \frac{1}{\tilde{c}} \sum_{i: s_i = H} w_i\right), \quad (8)$$

where $\tilde{c} = \sum_{i \in S} w_i / N \approx c$. In the equation (8), the sample size is enlarged from n to N based on the survey weights. Expression (8) takes into account uncertainty in the adjusted weights in mixture component allocation. One approach for bias adjustment is to replace the h th mixture weight λ_h by $\sum_{i \in S_h} w_i / (cN)$ in (4) in Theorem 1. This implies that selected subjects

in the data fully reflect information of the superpopulation density f_0 . However, for certain mixture components in f_0 , there may be very few or even zero subjects selected into the sample. This clearly leads to substantial uncertainty in the adjustment. To accommodate this uncertainty, we generate the adjusted weight $\tilde{\lambda}$ from (8), with the prior sample size for the different mixture components designed to represent uncertainty about the unsampled subjects. Posterior computation is straightforward: we simply apply any existing Markov chain Monte Carlo algorithm for mixture models to the selected sample, add sampling step (8) for generating the adjusted weights $\tilde{\lambda}$, and apply this adjustment to each step of the sampling algorithm to obtain samples from an adjusted posterior for the population density $f_0(y)$. As a default, we set $a_h = a$ for $h = 1, \dots, H$, with prior sample size $Ha \sim 1 - 2\%$ of population size N .

The proposed method can be applied to more broad cases. For example, if survey weights are available, we can construct strata after surveys based on the unique values of the weights (Si et al. (2015)). In this poststratification, the population can be divided into mutually exclusive subpopulations, each of which has a different inclusion probability, corresponding to the framework discussed above. Also, the population mean, median and standard deviation of y may be a major interest for practitioners in survey sampling, which can be easily estimated from the proposed adjusted density (7) and (8). Moreover, since the proposed method allows y to be any type of variable, it can be, for example, a time series $y = (y_1, \dots, y_T)'$ or a mixed-scale variable $y = (y_1, \dots, y_p)'$ where y_j has a different type of marginal distribution for $j = 1, \dots, p$.

3. Simulation Study

We illustrate performance of the proposed approach and compare to competitors. We consider a population with $N = 1,000,000$ consisting of three subpopulations having $N_1 = 650,000$, $N_2 = 300,000$ and $N_3 = 50,000$ with $v_m = N_m/N$. From each stratum, we randomly generate $n_m = 500$ subjects and construct survey weights by $w_i = N_m/n_m$ for $i \in D_m$ for $m = 1, 2, 3$. As competitors, we employ three model-based Bayesian methods. First, we consider a model-based Horvitz-Thompson estimator (Horvitz and Thompson (1952); Little (2004)), $y_i = \beta\pi_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \pi_i^2\sigma^2)$ where $\pi_i = 1/w_i$. Second, we consider a polynomial regression with random effects, $y_i = \beta_0 + \beta_1\pi_i + \beta_2\pi_i^2 + \gamma_{[i]} + \varepsilon_i$, $\varepsilon_i \sim N(0, \pi_i^2\sigma^2)$, $\gamma_m \sim N(0, \tau^2)$ where $\gamma_{[i]}$ denotes a random effect for the subpopulation to which the i th subject belongs. This can be induced by the spline model of Zheng and Little (2003). Also, we apply the Gaussian process regression model from Si et al. (2015), $y_i = \mu(x_{[i]}) + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, $\mu(x) \sim GP(\beta x, C)$, $C\{\mu(x_m), \mu(x_{m'})\} = \text{cov}\{\mu(x_m), \mu(x_{m'})\} = \tau^2 \exp(-\kappa|x_m - x_{m'}|)$ where $x_m = \log(w_m^*)$ and $x_{[i]}$ denotes the log weight for the stratum for the i th subject. In these approaches, the population density can be expressed as $f(y_i) = \sum_{w \in W} \pi(w) f(y_i|w)$ where $\pi(w)$ is the density of the survey weight with W the support of w . However, for simplicity in modeling, additional information of the true $\pi(w)$ is given to the competitors while it is uncertain for the proposed method. We also apply Dirichlet process mixtures without weight adjustment.

In the first case, we assume $f_1(y) = f_N(y | 2, 0.6)$, $f_2(y) = f_N(y | 0, 0.4)$ and $f_3(y) = f_N(y | -2, 0.3)$ in (1) where $f_N(y | a, b)$ denotes a normal density with mean a and standard deviation b . The mean and standard deviation decrease with the decreasing survey weight. For the proposed method, we use the Dirichlet process mixture of normals,

$$f_B(y) = \sum_{h=1}^H \lambda_h f_N(y | \mu_h, \tau_h) \text{ where } \lambda_h = V_h(1 - V_H), V_h \sim \text{Be}(1, \alpha), V_H = 1 \text{ with } H = 20.$$

Relying on an empirical Bayes method (Carlin and Louis (2000); Hoff (2009)), we use a conjugate prior using empirical information, $\alpha \sim \text{Ga}(0.25, 0.25)$, $\mu_h \sim N(\bar{y}, s_y^2)$, $\tau_h^2 \sim \text{Inverse-Gamma}(2, s_y^2/2)$ where $\bar{y}$ and s_y^2 are the sample mean and variance. This specification with a small prior sample size allows the data to dominate the prior effect (Dunson and Xing (2009); Si and Reiter (2013); Schifeling and Reiter (2015)). As for the prior in the step (8), we set $a_h = 1,000$ for each h . For competitors, we assume the empirical Bayes priors: $\beta \sim N(0, s_y^2)$, $\beta_j \sim N(0, s_y^2)$, $\sigma^2 \sim \text{Inverse-Gamma}(2, s_y^2/2)$, $\tau^2 \sim \text{Inverse-Gamma}(2, s_y^2/2)$ and $\kappa \sim \text{Ga}(1, 2)$. We draw 10,000 samples after the initial 5,000 samples are discarded as a burn-in period and every 10th sample is saved. Rates of convergence and mixing were adequate. Figure 1 shows the estimation results for case 1. The Horvitz-Thompson estimator fails to capture the multimodality, while the non-adjusted estimator has considerable bias. The random effect model and Gaussian process have somewhat better performance, but clear bias remains. Figure 3 reports the logarithm of KL divergence between the true and estimated densities. The proposed method accurately estimates the density with the smallest KL divergence.

We also considered a more complex density for each stratum, $f_1(y) = 0.2f_N(y | -2, 1) + 0.8f_N(y | 2, 0.8)$, $f_2(y) = 0.4f_N(y | -2, 1) + 0.6f_N(y | 2, 0.8)$ and $f_3(y) = 0.85f_N(y | -2, 1) + 0.15f_N(y | 2, 0.8)$. The Markov chain Monte Carlo settings are the same as in case 1. Figure 2 reports the result for case 2. The Horvitz-Thompson estimator, random effect model and Gaussian process regression work poorly, missing the multimodal shape of the true density because they construct population densities relying on unimodal densities for subpopulations. The non-adjusted method captures the bimodality but with substantial bias. The proposed method reports the smallest KL divergence in Figure 3.

Next, to assess the impact of increasing the number of strata while decreasing within-strata sample size, we consider a case with $M = 100$ in which $N_m = 1000m$, $n_m = 20$ for $m = 1, \dots, 100$ with $N = 5,050,000$ and $n = 2,000$ and $f_m(y) = f_N(y | -2, 0.3)$ for $m = 1, \dots, 30$, $f_m(y) = f_N(y | 0, 0.4)$ for $m = 31, \dots, 70$ and $f_m(y) = f_N(y | 2, 0.6)$ for $m = 71, \dots, 100$. We obtain a similar result to case 1 with the proposed method dominating competitors. To save space, we do not report the density estimation result. Figure 3 clearly shows the proposed method outperforms in terms of KL divergence.

Finally, we check results when the extent of dependence between the outcome and survey weights changes. Using the same setting as in the first two cases with three subpopulations, we assume y_i is generated from

$$y_i \sim N(q\mu(w_i), q\sigma^2(w_i) + 0.1), \quad \mu(w_i) = \sigma^2(w_i) = 0.01w_i.$$

Figure 4 shows the boxplots of the logarithm of KL divergence between the true and estimated densities with $q = 0, 0.5, 5$. In the case with $q = 0$ where the generating function is independent from the survey weights, the proposed method, non-adjusted method and Gaussian processes outperform the Horvitz-Thompson and polynomial regression probably because these two methods treat the variance as varying with the survey weights. Then, in the case with $q = 0.5$, the proposed method and Gaussian processes work similarly better than the other methods but, when the outcome becomes more correlated with the survey weights with $q = 5$, the proposed method shows the smallest log-KL difference.

4. Application to Adolescent Behaviour Analysis

We apply the proposed method to the National Longitudinal Study of Adolescent Health. Our focus is on studying the total number of sex partners in adolescence. The target population is adolescents in grades 7–12 in the United States during the 1994–95 school year with $N = 14,677,347$. The study design selects a stratified sample of high schools and middle schools from the U.S. with unequal probability of selection based on region, urbanicity, school type, ethnic mix, and size. Then, students in each school are stratified by grade and sex. In addition, the study drew supplemental samples, oversampling groups of particular interest based on ethnicity, genetic relatedness to siblings, adoption status, disability and parental education. Finally, survey weights have been constructed for population-representative inferences via weighting adjustments. The full study design is described by Harris et al. (2009). We use three waves of surveys in which participants are in grades 7–12 (1994–1995), young adults age 18–26 (2001–2002) and adults age 24–32 (2007–2008). In each wave of the public-use data, numbers of observations are 6447, 4812 and 4819, respectively. However, count data from sample surveys very seldom satisfy a parametric assumption, such as Poisson or negative binomial. As a flexible alternative, we apply the rounded kernel method in Canale and Dunson (2011) where latent continuous variables are modeled by (7) with Dirichlet process mixtures. Since we expect high right skew in these data, we use log cut-points instead of nonnegative integers, so that the Dirichlet process mixtures can efficiently approximate such distributions. For the priors of the latent continuous variable, we use $\mu_h \sim N(\tilde{y}, \tilde{s}_y^2)$, $\tau_h^2 \sim \text{Inverse-Gamma}(2, \tilde{s}_y^2/200)$ where $\tilde{y}$ and $\tilde{s}_y^2$ are the sample mean and variance of $\log(y_i + 0.5)$. Also, we set $a_h = 10,000$ for the Dirichlet prior in (8). We draw 20,000 samples after the initial 5,000 samples are discarded as a burn-in period and every 10th sample is saved. We observe that the sample paths were stable and the sample autocorrelations dropped smoothly.

Figure 5 shows the estimated probabilities for the three waves. 1994–1995 shows a high probability on zero with small values for positive counts. 2001–2002 expresses differences from 1994–1995 in that the probability on zero considerably decreases, while one shows the highest value and the tail gets heavy. The shape in 2007–2008 is similar to 2001–2002 in that both have highest probabilities at one and then steep declines. 2007–2008 shows a heavier tail with relatively high spikes at multiples of five. This is probably because people with many partners do not remember the exact numbers.

Acknowledgement

This work was supported by Nakajima Foundation and grants from the National Institutes of Health. The data are from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. The data are available at <http://www.cpc.unc.edu/addhealth>.

References

- Bellhouse DR, Stafford JE, 1999 Density estimation from complex surveys. *Statistica Sinica* 9, 407–424.
- Buskirk TD, 1998 Nonparametric density estimation using complex survey data, in: *Proceedings of the Survey Research Methods Section, American Statistical Association, Washington, DC* pp. 799–801.
- Buskirk TD, Lohr SL, 2005 Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference* 128, 165–190.
- Canale A, Dunson DB, 2011 Bayesian kernel mixtures for counts. *Journal of the American Statistical Association* 106, 1528–1539. [PubMed: 22523437]
- Carlin BP, Louis TA, 2000 *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall, London.
- Chen Q, Elliott MR, Little RJA, 2010 Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology* 36, 23–34. [PubMed: 29200606]
- Dunson DB, Xing C, 2009 Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* 104, 1042–1051.
- Gelman A, 2007 Struggles with survey weighting and regression modeling. *Statistical Science* 22, 153–164.
- Harris KM, Halpern CT, Whitsel E, Hussey J, Tabor J, Entzel P, Udry JR, 2009 The national longitudinal study of adolescent health: Research design [www document] URL: <http://www.cpc.unc.edu/projects/addhealth/design>.
- Hoff PD, 2009 *A first course in Bayesian statistical methods*. Springer, New York.
- Horvitz DG, Thompson DJ, 1952 A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association* 47, 663–685.
- Ishwaran H, James LF, 2001 Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 96, 161–173.
- Little RJA, 2004 To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* 99, 546–556.
- Schifeling TA, Reiter JP, 2015 Incorporating marginal prior information in latent class models. *Bayesian Analysis*, 1–20.
- Si Y, Pillai N, Gelman A, 2015 Bayesian nonparametric weighted sampling inference. *Bayesian Analysis* 10, 605–625.
- Si Y, Reiter JP, 2013 Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics* 38, 499–521.
- Zangeneh SZ, Little RJ, 2012 Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Proceedings of the Joint Statistical Meetings*.
- Zheng H, Little RJA, 2003 Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics* 19, 99–107.
- Zheng H, Little RJA, 2005 Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics* 21, 1–20.

Figure 1:

Estimated densities in case 1. Blue lines are the true density, red lines the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Nonadjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

Figure 2:

Estimated densities in case 2. Blue lines are the true density, red lines the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Non-adjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

Figure 3:

Boxplots of the logarithm of KL divergence between the true and estimated densities in Case 1 (left), Case 2 (middle) and Case 3 (right). PR is the proposed method, NA the non-adjusted Dirichlet process mixtures, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

Figure 4:

Boxplots of the logarithm of KL divergence between the true and estimated densities with $q = 0$ (left), 0.5 (middle) and 5 (right) in Case 4. PR is the proposed method, NA the non-adjusted Dirichlet process mixtures, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

Figure 5:

Estimated probabilities of total numbers of sex partners. The first row shows estimated probabilities in 1994-1995 (left), 2001-2002 (middle) and 2007-2008 (right). Lines with symbols show posterior means and dash lines 95% credible intervals. The second row shows posterior means of 0-5 partners (left), 6-15 (middle) and 15-40 (left). Red lines with circle represent posterior means for 1994-1995, blue lines with triangles for 2001-2002 and purple lines with squares for 2007-2008.