Bayesian Penalized Spline Model-Based Estimation of the Finite Population Proportion From Probability-Proportional-to-Size Samples

Qixuan Chen, Michael Elliott, Roderick Little
University of Michigan

Abstract

This paper develops Bayesian penalized spline predictive (BPSP) estimator of finite population proportion for probability-proportional-to-size samples. This new method allows the probabilities of inclusion to be directly incorporated into the estimation of population proportion, using a probit regression of the binary outcome on the penalized spline of the inclusion probabilities. The posterior distribution of the population proportion is then obtained using Gibbs sampling. Simulation studies show that the BPSP estimator gains efficiency over the HT estimator by using the inclusion probabilities in the non-sampled units, and that the BPSP estimator has a better coverage with narrower credible interval over the HT estimator, especially when the true population proportion is close to zero or one for small samples.

KEY WORDS: finite population proportion, Gibbs sampling, penalized spline regression model

1. Introduction

Suppose that we have a finite population consisting of *N* identifiable units and let binary variable *Y* be the characteristic of interest. In sample surveys, one of the most important problems is to estimate the proportion of units in the population with *Y* = 1.

$$p = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{1}$$

In probability-proportional-to-size (PPS) sampling, the inclusion probability π_i for unit *i* is proportional to the value x_i of size variable *X*, which is usually known for all units in the finite population before a sample is drawn. A PPS random sample *s* with elements $y_1,...y_n$ is drawn from the finite population according to the inclusion probabilities $\pi_1,...,\pi_N$.

A simple design-based estimator for *p* is the Horvitz-Thompson (1952) estimator, which weights cases by the inverse of their inclusion probabilities,

$$\hat{p}_{HT} = \frac{\sum_{i\in s} y_i/\pi_i}{\sum_{i\in s} 1/\pi_i} \tag{2}$$

Using Delta method and the Yates-Grundy (1953) estimator, by plugging in the joint inclusion probability approximation given by Hartley and Rao (1962), an estimated variance for $\hat{p}_{HT}$ is,

$$\hat{v}(\hat{p}_{HT}) = \frac{1}{N^2}\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\sum_{i<j}^{n}\left(\left(1-\pi_i-\pi_j+\sum_{k=1}^{N}\frac{\pi_k^2}{n}\right)\times\left(\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)-\hat{p}_{HT}\left(\frac{1}{\pi_i}-\frac{1}{\pi_j}\right)\right)^2\right)\right).$$

And an approximately $1-\alpha$ level confidence interval for the population proportion $\hat{p}_{HT}$ is given by

$$\left\{\hat{p}_{HT} - Z_{\alpha/2}[\hat{v}(\hat{p}_{HT})]^{1/2},\quad \hat{p}_{HT} + Z_{\alpha/2}[\hat{v}(\hat{p}_{HT})]^{1/2}\right\} \tag{3}$$

where $Z_{\alpha/2}$ is the $1-\alpha/2$ percentile of a standard normal distribution.

One limitation with the Horvitz-Thompson (HT) estimator is that it only incorporates the selection probabilities of units in the sample. But in PPS sampling, not only the inclusion probabilities in the sample but also those in the non-sampled units are known. Therefore, predictive estimators for statistical models relating y_i and π_i may improve the efficiency in estimating population proportion in a PPS sample.

On the other hand, the model-based predictive estimator may yield biased estimate when the underlying model is misspecified. This limitation motivates the development of flexible statistical models that are more robust to model misspecification. For continuous survey data, Zheng and Little (2003) estimated the finite population total using a nonparametric regression on the penalized spline (p-spline) of the inclusion probabilities. The estimator was shown to generally outperform the HT estimator and the generalized regression (GR) estimator in terms of root mean squared error. Zheng and Little (2005) also showed that p-spline model-based estimators and their jackknife standard errors lead to inferences that are superior to HT or GR based inferences.

The main purpose of the present work is to obtain a new estimator of the proportion of units in the finite population with *Y* = 1 from a PPS sample. First, the binary outcome *Y* is fitted on a p-spline of its inclusion probability by a probit regression in the sample. Then for the non-sampled units, *Y* is predicted based on the regression model and its corresponding inclusion probability. This model-based estimator is called Bayesian p-spline predictive (BPSP) estimator. The advantages of the BPSP estimator over the HT estimator are demonstrated by simulation studies.

2. Bayesian P-Spline Predictive (BPSP) Estimator

To understand the relationship between the binary outcome *Y* and the continuous inclusion probability π, we need to fit a binary regression of *Y* on π. Parametric binary regressions, such as the linear or quadratic logistic or probit model, may not adequate in fitting the data. One solution for this problem of inflexibility is to fit a binary regression on a spline of π by adding some knots. However, too many knots may result in the roughness of model fit. One way to overcome this problem is to retain all of the knots but to constrain their influence. This is called binary p-spline regression model.

Let $\Phi^{-1}(.)$ denote the inverse CDF of a standard normal distribution. We consider the following probit polynomial spline model with *m* truncated power bases (Ruppert, Wand, and Carroll 2003):

$$\Phi^{-1}\left(E\left(y_i \mid \beta,u,\pi_i\right)\right)=\beta_0+\sum_{k=1}^{p}\beta_k\pi_i^k+\sum_{l=1}^{m}b_l\left(\pi_i-k_l\right)_+^p$$
$$b_l \sim N\left(0,\tau^2\right) \qquad (4)$$
$$l=1,...,m\,;\,i=1,...,n\,.$$

The constants $k_1<...<k_m$ are *m* selected fixed knots. A function such as $(\pi_i-k)_+^p$ is called a truncated polynomial spline basis function with power *p*, and $(u)_+^p$ is equal to $\{u\times I(u\ge 0)\}^p$ for any real number *u*. Since the truncated polynomial spline basis function has *p*-1 continuous derivatives, higher values of *p* lead to smoother spline functions. In addition, by specifying a normal distribution for *u*, the influence of the *m* knots is constrained in Model (4).

Model (4) can also be written in the matrix form,

$$\Phi^{-1}\left(E\left(y_i \mid \beta,u,X,Z\right)\right)=\left(X\beta+Zb\right)_i,\ i=1,...,n$$
$$\beta=\left(\beta_1,...,\beta_p\right)^T,\ b=\left(b_1,...,b_m\right)\sim N_m\left(0,\tau^2 I_m\right)$$
$$X=\begin{pmatrix}1 & \pi_1 & ... & \pi_1^p\\ . & . & ... & .\\ . & . & ... & .\\ . & . & ... & .\\ 1 & \pi_n & ... & \pi_n^p\end{pmatrix},Z=\begin{pmatrix}(\pi_1-k_1)_+^p & ... & (\pi_1-k_m)_+^p\\ . & ... & .\\ . & ... & .\\ . & ... & .\\ (\pi_n-k_1)_+^p & ... & (\pi_n-k_m)_+^p\end{pmatrix}.$$

And it can be implemented using the generalized linear mixed models. However, their fitting presents some computational challenges. Alternatively, using the idea of data augmentation, Gibbs sampling makes the computation much simpler (Ruppert, Wand, and Carroll 2003). The working algorithm is as follow:

a) Probit regression models have computational advantages compared to logistic regression models. The probit regression model for the binary outcome $y=[y_1,...,y_n]^T$ corresponds to a normal regression model for a latent continuous data $y^*=\left[y_1^*,...,y_n^*\right]^T$, which has a truncated multivariate normal distribution with mean $(X\beta+Zb)$ and identity covariance matrix (Albert and Chib 1993). The relation to the observed binary response data y_i is that y_i is the indicator that $y_i^*>0$. With some initial values of (β,b), values of the latent continuous data y_i^* can be simulated.

b) Specifying an improper uniform prior on β and an inverse gamma $IG(0.01,0.01)$ on τ^2, the posterior distribution of $\left(\beta,b,\tau^2\right)$ given the simulated latent continuous data y^* is

$$(\beta,b)\mid\tau^2,y^*$$
$$\sim MVN_{m+p+1}\left(\left(C^TC+D/\tau^2\right)^{-1}C^Ty^*,\left(C^TC+D/\tau^2\right)^{-1}\right)$$
$$\tau^2\mid\beta,b\sim IG\left(0.01+m/2,0.01+\|b\|^2/2\right), \qquad (5)$$

where $C=[X, Z]$ and *D* is a diagonal matrix with $p+1$ zeros followed by *m* ones on the diagonal.

c) At iteration *t*, draws from the posterior distribution of $\left(\beta,b\mid y^{*(t-1)},\tau^{2(t-1)}\right)$ in Equation (5) are used to sample new latent data $\hat{y}^{*(t)}$ conditional on observed binary variable *y* for the sample, and to obtain the predicted values $\hat{y}^{(t)}$ as indicators whether the sampled $\hat{y}^{*(t)}$ are greater than zero for non-sample units. And then the predictive proportion is calculated as

$$\hat{p}^{(t)}=N^{-1}\sum_{i\in s}y_i+N^{-1}\sum_{j\notin s}\hat{y}_j^{(t)} \qquad (6)$$

d) The draws of $\hat{y}^{*(t)}$ in the sample are used to draw a new $\left(\beta,b,\tau^2\right)$ at iteration $t+1$. The posterior distribution of $\hat{p}$ is then obtained by the above Gibbs sampler.

The posterior mean of $\hat{p}$ is called the Bayesian p-spline predictive (BPSP) estimator of the finite population proportion, and is denoted as $\hat{p}_{BPSP}$. One advantage of Bayesian analysis is that the whole posterior distribution of the parameter of interest is available. To keep the distinction between Bayesian and classical inference clear, the Bayesian intervals are referred to as credible intervals rather than confidence intervals. The $1-\alpha$ level credible interval for the BPSP estimator of population proportion can be formed in many different ways. One simple way is to split the α equally between the upper and lower endpoints, and thus the $1-\alpha$ credible interval is $\{p:\inf\{q_L:\hat{p}_{BPSP}(q_L)\ge\alpha/2\}\le p\le\inf\{q_U:\hat{p}_{BPSP}(q_U)\ge 1-\alpha/2\}\}$.

3. Simulation Study

3.1 Design of the simulation studies

Simulation studies are conducted to study the performance of the BPSP estimator, compared to the HT estimator for four different kinds of populations with two different sampling rates.

We simulate finite populations with a population size of 2000. The inclusion probabilities in the population, $\pi = \{\pi_1, \ldots, \pi_{2000}\}$, are simulated as proportional to the consecutive positive integer values: 71, 72, ..., 2070. Continuous data $Z = \{z_1, \ldots, z_{2000}\}$ are generated from normal distributions with means structures $f(\pi)$ and constant error variance 0.04. The binary outcomes $Y = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$ are then created by using the superpopulation 10th, 25th, 50th, 75th, and 90th percentiles of *Z* as cut-off values. For instance, Y_1 equals to 1 if *Z* is less than its superpopulation 10th percentile, otherwise 0. The target of estimation is the population proportion with *Y* equal to 1. Four different mean structures of the inclusion probabilities are simulated for *Z*. They are constant function (NULL) $f(\pi_i) \equiv 0.30$, linearly increasing function (LINUP) $f(\pi_i) = k_1 \pi_i$, linearly decreasing function (LINDOWN) $f(\pi_i) = k_2 \pi_i$ and exponentially increasing function (EXP) $f(\pi_i) = \exp(-4.64 + k_3 \pi_i)$.

We simulate two different sample sizes 100 and 200 from each mean structure of population. In each replicate, a finite population is generated before a sample is drawn. A PPS sample is then drawn systematically from a randomly ordered list of the finite population. For each population and sample size combination, 1000 replicates are obtained and the two estimators are compared in terms of bias, root mean square error (RMSE), length of the 95% confidence /credible interval, and the coverage.

Model (4) is specified using linear, quadratic, or cubic splines, and is also fitted using various numbers and positioning of knots: with 5, 15, or 30 knots, and placing the knots in equal space or in the equally spaced sample percentiles of the inclusion probabilities. Simulations suggest that linear splines perform as well as quadratic splines or cubic splines in this setting. In addition, simulations suggest that the number and positioning of the knots do not make much difference to the results, provided that the knots cover the range of the data in the sample reasonably well. Therefore, we only present the simulation results for the linear BPSP estimator with knots placed at the 15 evenly spaced sample percentiles. Figures 1-4 show how Bayesian probit regression models are fitted for binary outcomes on the p-spline of the inclusion probabilities in these four mean structures. Simulation results are shown in Tables 1 through 4.

3.2 Simulation results

Figures 1-4 show that the Bayesian p-spline probit regression models fit well for the binary outcomes. In each figure, the upper left plot is the scatter plot of the continuous variable *Z* by the inclusion probabilities, with five horizontal parallel lines superimposed, representing the superpopulation 10th, 25th, 50th, 75th, and 90th percentiles respectively. In the upper middle plot, the binary variable *Y*, defined as 1 if *Z* is less than the superpopulation 10th percentile, are plotted with black circles, and the true probabilities of *Y* = 1 over inclusion probabilities are plotted with a solid black curve. The solid grey curve and two dashed grey curves are the mean and 95% credible intervals of the posterior distribution of probabilities of *Y* = 1 given the inclusion probabilities, obtained from the Bayesian p-spline probit regression model. The other four plots are similar to the upper middle plot, but with superpopulation 25th, 50th, 75th, and 90th percentiles as cut-off values in defining *Y*. These plots show that the true probabilities of *Y* = 1 fall within the 95% credible intervals of the fitted values, and are close to the mean fitted values. Similar results are found in Figure 2, 3 and 4.

Tables 1-4 show that the BPSP estimator outperforms the HT estimator in terms of RMSE for the finite population proportion estimation, though the BPSP estimator is slightly more biased than the HT estimator. In the "EXP" case, when the true value of population proportion is 0.75 or 0.90, the RMSE is reduced by about a half by using the p-spline model-based predictive estimator. This means that the BPSP estimator is more efficient than the HT estimator and gains more when the binary outcome is better separated by the probabilities of inclusion.

In general, the coverage calculated based on the BPSP estimator and its credible interval is closer to its nominal level than the HT estimator and its confidence interval, especially when the population proportion is closer to 0 or 1, and there are relatively few observations in the tails. For example, for the "NULL" case, only about 85% and 90% of the 95% confidence intervals of the HT estimator cover the true value of population proportion with sample sizes 100 and 200 respectively; while the coverages are closer to 95% for the BPSP estimator. Similarly, we see the big improvement in the coverages of the credible intervals of the BPSP estimator over the HT estimator when the true population proportion is close to 0 in both "LINUP" and "EXP" cases, and when the true population proportion is close to 1 in "LINDOWN" case, since the probabilities of inclusion are smaller for the lower tails of "LINUP" and "EXP" and for the upper tail of "LINDOWN", which results in fewer data are selected into the sample in these groups. More important, we achieve narrower intervals with the BPSP estimator than the HT estimator while obtaining better coverages.

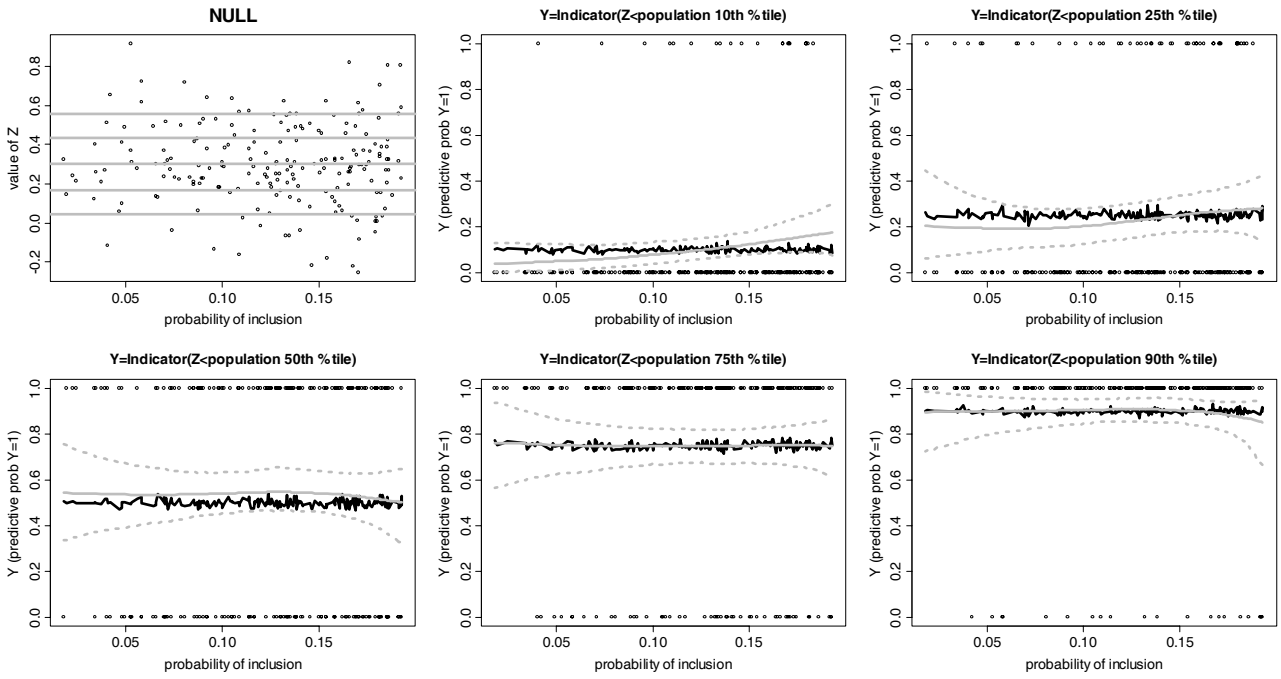

Figure 1: "NULL" population (n=200, N=2000)

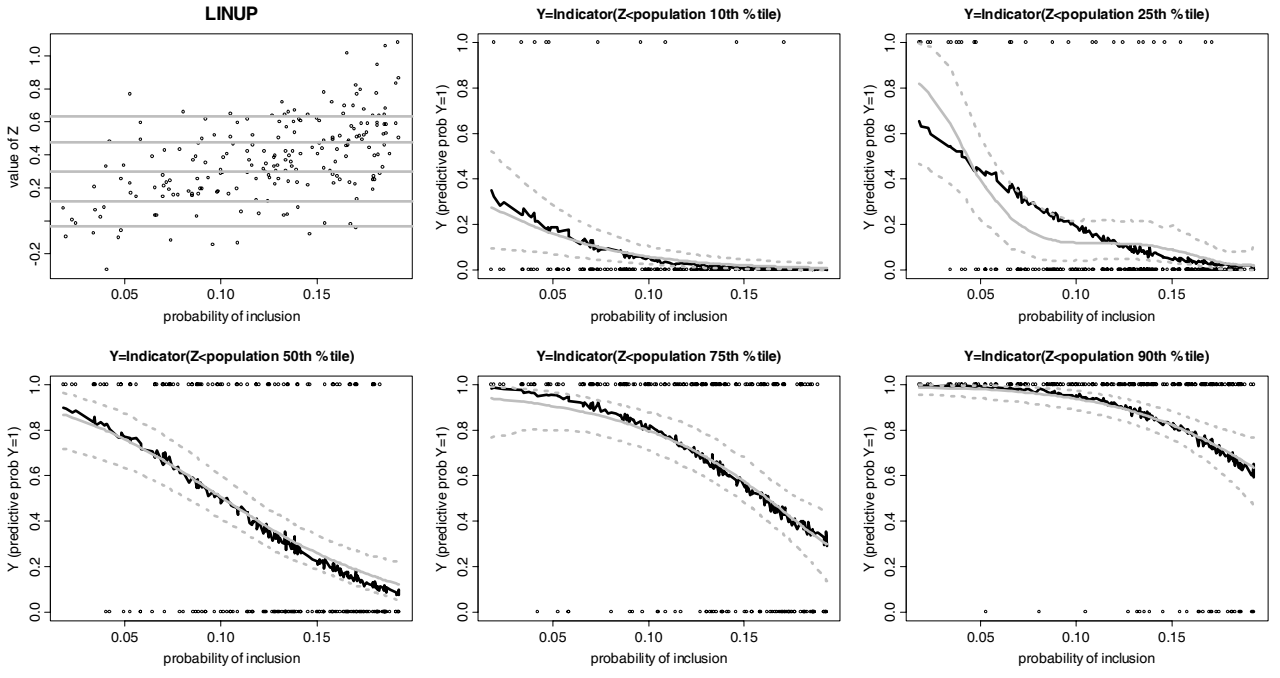

Figure 2: "LINUP" population (n=200, N=2000)

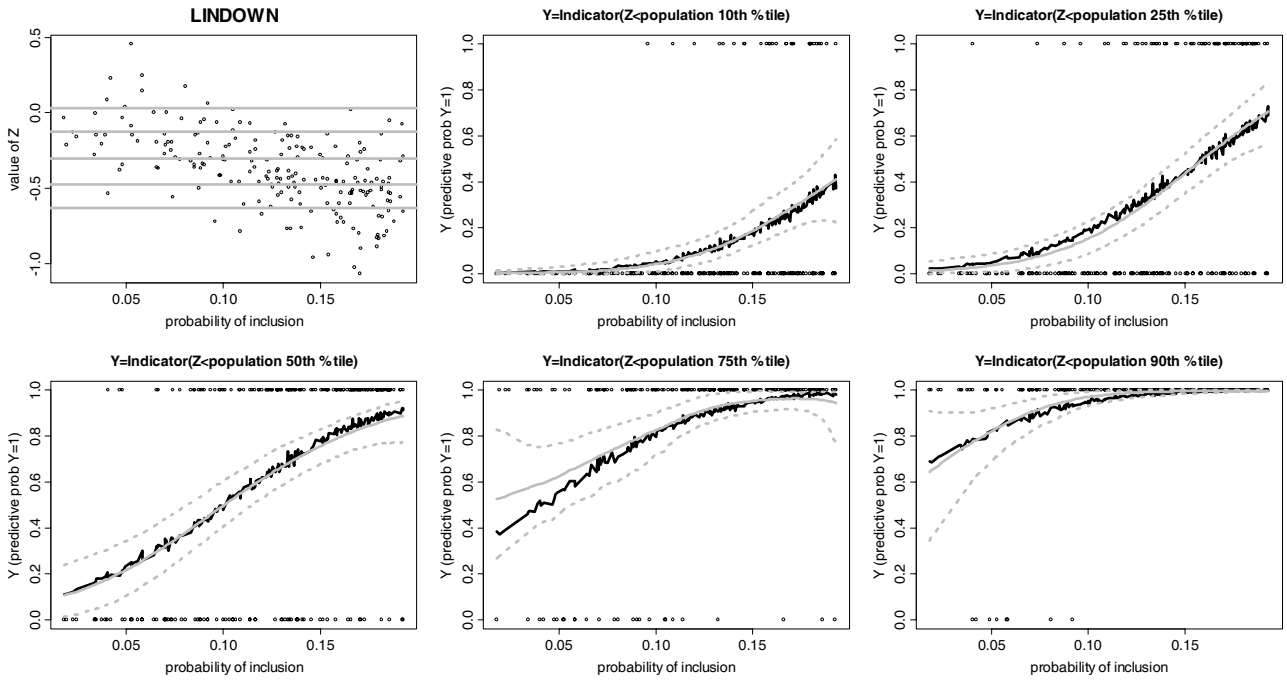

Figure 3: "LINDOWN" population (n=200, N=2000)

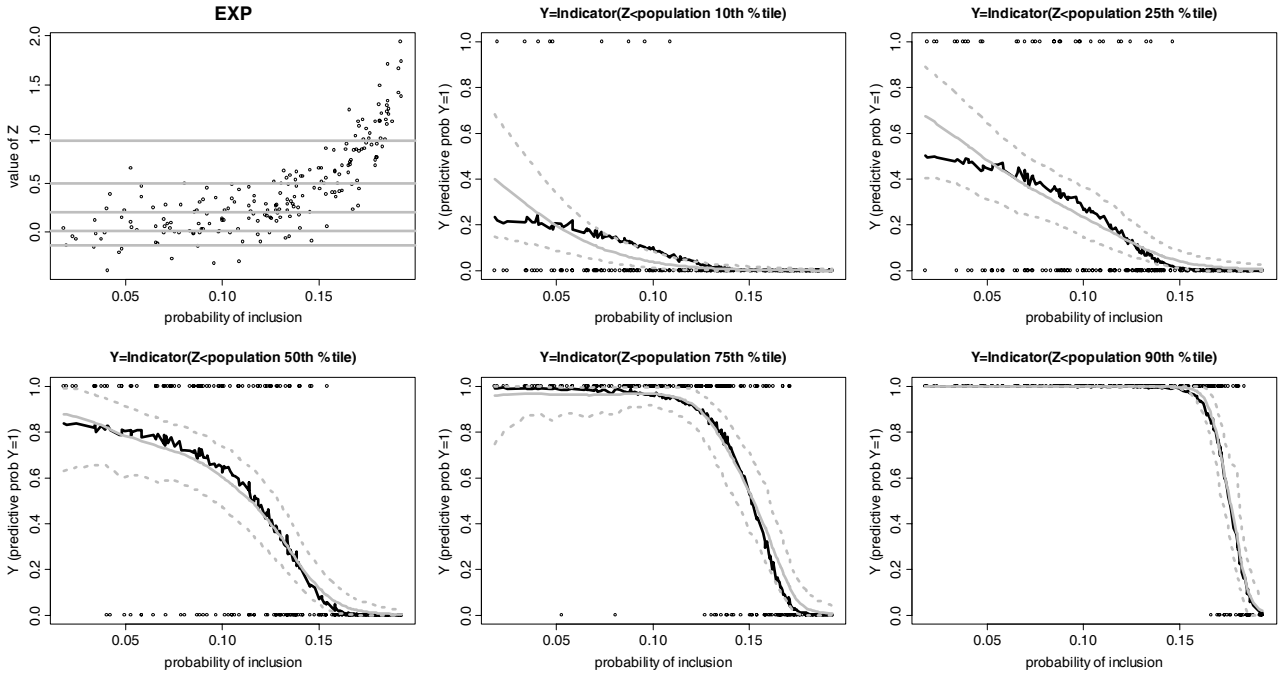

Figure 4: "EXP" population (n=200, N=2000)

Table 1: "NULL" population $f(\pi_i) \equiv 0.30$

Sample size	True proportion	bias*1000		RMSE*1000		length*100		coverage*100	
		HT	BPSP	HT	BPSP	HT	BPSP	HT	BPSP
100	0.10	-0.7	5.5	40.6	38.1	13.9	13.6	84.6	92.7
	0.25	0.9	3.8	58.4	52.6	21.6	19.3	90.4	93.0
	0.50	-0.7	-1.2	65.5	58.6	25.1	22.1	93.1	94.7
	0.75	-2.4	-6.3	59.5	54.3	21.5	19.3	91.1	93.2
	0.90	-0.7	-6.7	42.0	39.3	14.1	13.7	85.2	90.6
200	0.10	0.1	3.2	27.9	25.7	10.0	9.6	90.6	92.3
	0.25	-1.3	1.5	40.4	36.7	15.0	13.7	92.0	94.4
	0.50	-1.1	-0.2	48.1	43.2	17.7	15.8	92.0	93.1
	0.75	0.3	-2.2	41.3	37.8	15.1	13.7	90.7	93.6
	0.90	-0.1	-3.1	28.6	26.7	10.2	9.7	89.2	93.8

Table 2: "LINUP" population $f(\pi_i) = k_1\pi_i$

Sample size	True proportion	bias*1000		RMSE*1000		length*100		coverage*100	
		HT	BPSP	HT	BPSP	HT	BPSP	HT	BPSP
K_1=6 100	0.10	-0.01	8.2	55.5	47.6	19.4	15.5	83.6	90.7
	0.25	-3.0	-0.6	71.4	55.0	26.2	19.3	89.6	93.2
	0.50	-4.0	-5.3	66.1	48.9	25.3	18.4	91.7	94.3
	0.75	-1.8	-4.0	47.5	37.7	17.4	13.8	91.5	94.0
	0.90	-0.4	-2.9	27.3	24.2	10.2	9.2	91.3	92.9
K_1=3 200	0.10	-2.5	3.4	39.1	32.4	14.1	11.8	86.0	94.3
	0.25	-1.7	0.9	48.7	37.9	18.7	14.0	92.0	94.5
	0.50	-1.4	-1.4	47.8	35.1	17.7	12.8	92.5	92.1
	0.75	-0.1	-1.4	33.2	26.2	12.0	9.4	93.1	93.1
	0.90	0.7	-0.7	19.2	17.1	6.9	9.1	92.3	92.7

Table 3: "LINDOWN" population $f(\pi_i) = -k_2\pi_i$

Sample size	True proportion	bias*1000		RMSE*1000		length*100		coverage*100	
		HT	BPSP	HT	BPSP	HT	BPSP	HT	BPSP
K_2=6 100	0.10	-0.4	2.0	26.8	23.9	10.1	9.1	91.3	93.6
	0.25	1.2	3.5	46.3	36.8	17.5	13.8	93.2	93.5
	0.50	-0.2	1.3	68.2	50.1	25.2	18.2	91.5	92.5
	0.75	-0.05	-3.2	70.2	52.4	26.3	19.3	90.1	94.0
	0.90	-1.0	-9.9	57.5	48.3	19.5	15.6	82.9	91.2
K_2=3 200	0.10	1.0	2.2	19.5	17.5	7.0	7.0	92.8	91.7
	0.25	0.5	1.4	33.0	26.0	12.1	12.1	93.3	93.0
	0.50	2.0	2.1	45.9	33.5	17.7	17.7	93.9	93.0
	0.75	1.8	0.1	49.3	37.8	18.9	18.9	92.2	93.9
	0.90	-0.5	-5.9	39.8	33.5	14.6	14.6	87.8	93.3

Table 4: "EXP" population $f(\pi_i) = \exp(-4.64 + k_3\pi_i)$

Sample size	True proportion	bias*1000		RMSE*1000		length*100		coverage*100	
		HT	BPSP	HT	BPSP	HT	BPSP	HT	BPSP
K_3=52 100	0.10	1.2	16.7	51.5	52.0	17.9	16.0	84.6	89.9
	0.25	-1.2	12.1	67.8	57.8	25.1	19.4	90.1	91.7
	0.50	-4.0	-2.1	66.9	48.2	25.3	16.8	91.2	91.5
	0.75	-2.0	-3.4	43.8	23.3	16.6	8.3	92.1	92.4
	0.90	-1.3	-1.1	24.6	13.0	9.4	4.6	93.0	90.3
K_3=26 200	0.10	-2.0	10.4	35.7	34.6	12.7	12.1	87.7	93.3
	0.25	-1.8	9.5	47.6	42.7	17.7	14.5	92.1	92.0
	0.50	-0.8	-0.1	47.6	34.5	17.7	12.4	92.3	91.6
	0.75	0.7	-1.5	29.4	16.2	11.4	5.8	93.7	92.0
	0.90	0.2	-0.5	16.5	8.6	6.3	3.2	94.6	92.2

5. Discussion

The Bayesian p-spline predictive estimator is preferable to the Horvitz-Thompson estimator in probability-proportional-to-size sampling, because the BPSP estimator gains efficiency over the HT estimator by using the inclusion probabilities in the non-sampled units, using a flexible statistical model that is robust to model misspecification. Though the bias is slightly increased in some cases, the gain in efficiency more than offsets the loss in bias. In addition, the coverage calculated based on the BPSP estimator and its credible interval is closer to its nominal level than the HT estimator and its confidence interval, especially when the population proportion is closer to 0 or 1, and there are relatively few data in the tails.

The HT estimator and its approximate 95% confidence interval can provide a valid inference for population proportion when the sample is large. However, when the sample is moderate or small and the true population proportion to be estimated is close to 0 or 1, the asymptotic properties do not hold for the HT estimator any more. However, Bayesian approach can provide more valid inference for small samples, and thus the BPSP estimator has a better coverage with narrower credible interval in these cases.

Compared to the HT estimator, one drawback of the BPSP estimator is that it needs more computation; but the extra computation provides the complete posterior distribution of the population proportion, that is, we don't need extra work to estimate the variance or 95% credible interval of the BPSP estimator, as they can be obtained simultaneously with the point estimator. In Zheng and Little (2005), three variance estimators of the p-spline model-based estimator for finite population total in a PPS sample were compared, including model-based empirical Bayes variance estimator, jackknife method of variance estimation, and the balanced repeated replicate (BRR) variance estimation method. The simulation studies showed that the jackknife method worked well, whereas the BRR method tended to yield conservative standard errors and the model-based empirical Bayes estimator was vulnerable to misspecification of the variance structure. In the present work, the $1-\alpha$ level credible interval for the BPSP estimator of population proportion is constructed by splitting α equally between the upper and lower endpoints of the posterior distribution of p. This pure Bayesian approach based on Gibbs sampling and draws from the posterior distributions can avoid specification of the variance structure and heavy computation associated with the jackknife and BRR method.

In future work we plan to extend the BPSP estimator of the proportion to include covariates other than inclusion probabilities, and from one-stage PPS sampling to two-stage PPS sampling with clusters. We also plan to adapt the current approach for finite population proportion to the related problem of estimating finite population percentiles.

Acknowledgment

This work is supported by The Dow Chemical Company through an unrestricted grant to the University of Michigan.

References

1. James H. Albert, Siddhartha Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of American Statistical Association.*, 88, 422, 669-679
2. Hartley, H.O. and Rao, J.N.K. (1962). "Sampling with Unequal Probabilities and without Replacement." *Annals of Mathematical Statistics*, 33, 350–374
3. Horvitz, D. G., and Thompson, M. E. (1952). "A generalization of sampling without replacement from a finite universe." *Journal of American Statistical Association.* 47, 663-685.
4. Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.
5. Yates, F., and Grundy, P. M. (1953). "Selection without replacement from within strata with probability proportional to size." *Journal of the Royal Statistical Society, Series B.* 15, 235-261.
6. Zheng, H. and Little, R.J.A. (2003). "Penalized Spline Model-Based Estimation of Finite Population Total from Probability-Proportional-to-Size Samples." *Journal of Official Statistics*, 19, 2, 99–117.
7. Zheng, H. and Little, R.J.A. (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, 1-20.