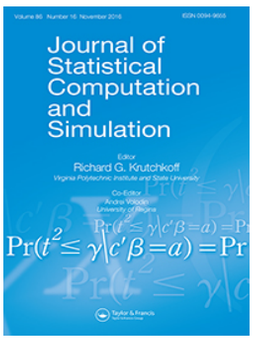

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

A nonparametric Bayesian prediction interval for a finite population mean

Balgobin Nandram & Jiani Yin

To cite this article: Balgobin Nandram & Jiani Yin (2016) A nonparametric Bayesian prediction interval for a finite population mean, Journal of Statistical Computation and Simulation, 86:16, 3141-3157, DOI: 10.1080/00949655.2016.1151518

To link to this article: http://dx.doi.org/10.1080/00949655.2016.1151518

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2016
VOL. 86, NO. 16, 3141–3157
http://dx.doi.org/10.1080/00949655.2016.1151518

Taylor & Francis
Taylor & Francis Group

A nonparametric Bayesian prediction interval for a finite population mean

Balgobin Nandram and Jiani Yin

Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, USA

ABSTRACT

Given a sample from a finite population, we provide a nonparametric Bayesian prediction interval for a finite population mean when a standard normal assumption may be tenuous. We will do so using a Dirichlet process (DP), a nonparametric Bayesian procedure which is currently receiving much attention. An asymptotic Bayesian prediction interval is well known but it does not incorporate all the features of the DP. We show how to compute the exact prediction interval under the full Bayesian DP model. However, under the DP, when the population size is much larger than the sample size, the computational task becomes expensive. Therefore, for simplicity one might still want to consider useful and accurate approximations to the prediction interval. For this purpose, we provide a Bayesian procedure which approximates the distribution using the exchangeability property (correlation) of the DP together with normality. We compare the exact interval and our approximate interval with three standard intervals, namely the design-based interval under simple random sampling, an empirical Bayes interval and a moment-based interval which uses the mean and variance under the DP. However, these latter three intervals do not fully utilize the posterior distribution of the finite population mean under the DP. Using several numerical examples and a simulation study we show that our approximate Bayesian interval is a good competitor to the exact Bayesian interval for different combinations of sample sizes and population sizes.

ARTICLE HISTORY
Received 31 December 2014
Accepted 2 February 2016

KEYWORDS
Correlation; exchangeability; nonparametric procedure; normality; Polya posterior; sampling-based method

1. Introduction

We assume that a simple random sample is drawn from a finite population and the population values come from a Dirichlet process (DP).[1] We provide a Bayesian prediction interval for the finite population mean. This problem can be solved easily. However, when the population size is much larger than the sample size, the computation becomes prohibitive. Thus, we obtain an approximate interval which is virtually the same as the exact Bayesian interval for large populations. This is obtained using the central limit theorem for exchangeable random variables. As competitors we also consider other approximations such as those based on the posterior mean (PM) and variance of the finite population mean together with the assumption of normality.

All survey data are inherently discrete (e.g. by truncation or rounding). We envision data which may come in latent clusters (i.e. gaps in the data) or data which have ties and a parametric distribution is unlikely to provide a good fit. This is particularly true for survey data. For example, income data are typically recorded in thousands of dollars and body mass index data are recorded in integers. Binder [2] and Lo [3] provided an asymptotic interval for the finite population mean under the DP. Lo [3]

CONTACT Balgobin Nandram balnan@wpi.edu

went a bit further to obtain large-sample Bayes confidence band for the finite population distribution in terms of the distribution function of the sample. In addition, Lo [4] discussed the Bayesian bootstrap for finite population with prediction intervals. Under stratification both Binder [2] and Lo [3] showed how the asymptotic interval for the random sampling extends easily to cover stratification as well.

It is apparent that statisticians are looking for nonparametric methods to obtain more robust procedures than are provided by parametric models such as those based on normality. For this reason, the purpose of our paper is to obtain a closed-form nonparametric interval estimator of a finite population mean when a sample is obtained from a relatively much larger population. One way to proceed is to use a DP model, an important model in nonparametric Bayesian statistics, which can provide a relatively robust interval estimator. Nonparametric Bayesian statistics is currently a very active area [5] and it is used in many applications. In this paper we report an interesting nonparametric prediction interval for a finite population mean under a DP when a simple random sample is available.

We have a simple random sample of size n from a population of size N. We assume that the sampled values are $y_1, \ldots, y_n$ and the nonsampled values are $y_{n+1}, \ldots, y_N$. Inference is required for the finite population mean, $\bar{Y} = \sum_{i=1}^{N} y_i/N$, and data $y_1, \ldots, y_n$ are available. Specifically, a Bayesian prediction interval is needed for $\bar{Y}$. We wish to carry out a nonparametric Bayesian analysis. The DP model is

$$y_1, \ldots, y_N \mid G \overset{\text{iid}}{\sim} G \quad \text{and} \quad G \mid \alpha, H_{\underset{\sim}{\theta}}(y) \sim \text{DP}\{\alpha, H_{\underset{\sim}{\theta}}(y)\},$$

where α is the concentration parameter and $H_{\underset{\sim}{\theta}}(y)$ is the baseline distribution corresponding to a parametric density function, and is therefore a function of unknown parameters $\underset{\sim}{\theta}$. In the DP α is a measure of the variability of the random distribution, G, around $H_{\underset{\sim}{\theta}}(y)$, a family of continuous distributions indexed by $\underset{\sim}{\theta}$. Motivated by Binder [2] and Lo [3] and using the DP model, we construct a nonparametric Bayesian prediction interval for $\bar{Y}$. With an effort to obtain a closed-form nonparametric Bayesian prediction interval, we take $H_{\underset{\sim}{\theta}}(y)$ to be the normal cumulative distribution function (i.e. $y \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$).

It is easy to show that $\text{Cor}(y_i, y_j) = 1/(1+\alpha)$, $i \neq j$. This correlation is useful because when there are gaps or ties in data, it is reasonable to believe that the data are correlated. It is difficult to model such a correlation unless one identifies the appropriate structure (e.g. clustering) in the data. But this seems problematic because we have a simple random sample. Of course, in the design-based analysis the sample indicators are negatively correlated under simple random sampling without replacement. But in any model-based analysis of simple random sampling the assumption that is usually used is independent and identical values.

In this paper we use the DP to model a simple random sample from a finite population. This requires generating the nonsampled finite population. But this can be prohibitive because when the population size is much larger than the sample size, the computation time is intolerable. This gives the exact prediction interval. (Throughout this paper we use the word 'exact' to refer to situations where no analytical approximations are used. A situation in which a sampling-based method is used is an example.) So we obtain a competitive approximation to the exact prediction interval.

The main reason for using the DP is to accommodate the gaps and ties in the data. The exchangeability and the gaps and ties in the responses appear contradictory. However, this is not true because we are not restricted to independent and identically distributed responses from a common parametric distribution but rather from a random distribution which follows a DP (Appendix 1). When the random distribution is integrated out, the responses become equi-correlated. Moreover, under the DP the responses are discrete with probability one (Appendix 1), thereby making the DP a natural clustering algorithm. Even when a simple random sample is taken from a population, there may be hidden structures in the data that the DP can accommodate because it is essentially nonparametric.

Our theoretical work consists of three theorems which are useful for different purposes. Theorem 1 is a statement about propriety of the joint posterior density under the DP model. This is useful because

if the posterior density is improper, inference about the finite population mean will be defective (i.e. the coverage of the prediction interval will be unknown). Thus, Theorem 1 adds credibility to the Bayesian procedure. Theorem 2 is a convenient statement of Binder's formulae for the PM and variance of the finite population mean conditional on the hyperparameters in the DP model. This is useful in the construction of our approximate Bayesian prediction interval. Theorem 3 uses Theorem 2 to obtain the PM and variance under the DP model when the hyperparameters are integrated out.

Our predictive interval may be useful in two situations for 'Big Data Science' today. First, a sample may be available from a much larger population and prediction is needed for the entire population (e.g. the finite population mean). Second, we actually have the data from the entire population. But because the computational effort needed to analyse the data are enormous, one can take a small sample. In general, it is difficult to tell what the actual data distribution is, and our nonparametric prediction interval works well in either case.

The plan of the paper is as follows. In Section 2 we first discuss the design-based prediction interval. Then we describe the exact Bayesian interval and the approximate Bayesian interval (ABI). We also describe two additional approximate intervals, empirical Bayes interval and the exact moment interval. It is convenient to present the three theorems as we proceed with the discussion on the Bayesian methodology. In Section 3 we explore 14 numerical examples and a simulation study to compare the five prediction intervals. In Section 4 we have concluding remarks. A review of the DP and proofs of two new theorems are given in the appendices.

2. Bayesian methodology

We have a simple random sample of size n from a finite population of size N. Let $y_1, \ldots, y_n$ denote the sampled (s) values and $y_{n+1}, \ldots, y_N$ denote the nonsampled (ns) values. We observe $\underset{\sim}{y}_s = (y_1, \ldots, y_n)$ but not $\underset{\sim}{y}_{ns} = (y_{n+1}, \ldots, y_N)$. We need a prediction interval for the finite population mean, $\bar{Y} = \sum_{i=1}^{N} y_i/N = f\bar{y}_s + (1-f)\bar{y}_{ns}$, where $f = n/N$ is the sampling fraction, $\bar{y} = \sum_{i=1}^{n} y_i/n$, the sample mean, and $\bar{y}_{ns} = \sum_{i=n+1}^{N} y_i/(N-n)$, the nonsample mean which is to be predicted. Also, let $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$ denote the sample variance.

First, we review a well-known prediction interval. Under simple random sampling a 95% prediction interval for $\bar{Y}$ is

$$\bar{y} \pm z_{2.5}\sqrt{\frac{1-f}{n}}s, \tag{1}$$

where $z_{2.5}$ is the 2.5th percentile point of the standard normal density. For large n the prediction interval in Equation (1) is an approximate (normality) 95% Bayesian prediction interval. We call this interval the design-based interval (DBI) and the design-based method (DBM), and it is pertinent to start with it. This is the asymptotic prediction interval obtained by Binder [2] and Lo [3] under the DP.

Note that if we assume the Bayesian model

$$y_1, \ldots, y_N \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2), \pi(\mu, \sigma^2) \propto 1/\sigma^2, \quad -\infty < \mu < \infty, \ \sigma^2 > 0,$$

the Bayesian prediction interval is

$$\bar{y} \pm t_{n-1,2.5}\sqrt{\frac{1-f}{n}}s,$$

where $t_{n-1,2.5}$ is the 2.5th percentile of the Student's t density on $n-1$ degrees of freedom. This is true because the prior predictive distribution of $\bar{Y}$ is normal with mean $f\bar{y} + (1-f)\mu$ and variance $(1-f)(\sigma^2/N)$, $\mu \mid \sigma^2, \bar{y} \sim N(\bar{y}, \sigma^2/n)$ and $(n-1)s^2/\sigma^2 \mid s^2 \sim \chi^2_{n-1}$. For a large n this latter interval is given in Equation (1). However, it is well known that this latter interval is not robust to non-normality especially when the sample size is small.

We use a DP model for the population values to construct a 95% nonparametric Bayesian prediction interval for a finite population mean. The DP is given by

$$y_1, \ldots, y_N \mid G \overset{\text{iid}}{\sim} G \quad \text{and} \quad G \mid \alpha, H_{\underset{\sim}{\theta}}(y) \sim \text{DP}\{\alpha, H_{\underset{\sim}{\theta}}(y)\}, \tag{2}$$

where the mean and variance of the DP are, respectively, given by $\text{E}\{G(y)\} = H_{\underset{\sim}{\theta}}(y)$ and $\text{Var}\{G(y)\} = H_{\underset{\sim}{\theta}}(y)\{1 - H_{\underset{\sim}{\theta}}(y)\}/(\alpha + 1)$. In this paper we take $H_{\underset{\sim}{\theta}}(y)$ to be

$$H_{\underset{\sim}{\theta}}(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} \text{e}^{-(1/2\sigma^2)(t-\mu)^2}\, \text{d}t, \quad -\infty < y < \infty,$$

the cdf of the normal random variable with mean μ and variance σ^2 (i.e. $\underset{\sim}{\theta} = (\mu, \sigma^2)$). Observe that $H_{\underset{\sim}{\theta}}(y)$ is fully parametric and there should be some care in its specification [6] even though this is not our current issue. Our objective is to obtain a 95% prediction interval for $\bar{Y}$ under the DP model. As mentioned, Appendix 1 provides a more detailed review about the DP.

In Section 2.1 we obtain a full Bayesian prediction interval for $\bar{Y}$ in which the hyperparameters have prior distributions. In Section 2.2 we obtain the approximate Bayesian prediction interval that is a serious competitor to the full Bayesian prediction interval. In Section 2.3, for comparison, we develop two additional approximate prediction intervals that are based on moments.

2.1. Full Bayesian prediction interval

In this section we describe the full Bayesian model and we show how to obtain the full Bayesian prediction interval.

The full Bayesian model is obtained by putting prior distributions on α, μ and σ^2. In this paper, we will use a 'Cauchy' type prior, sometimes called a shrinkage prior, of the following form for α, $p(\alpha) = 1/(\alpha + 1)^2, \alpha > 0$ (a f density with two degrees of freedom in both the numerator and denominator). It is slightly more convenient to use $p(\alpha) = 1/(\alpha + 1)^2, \alpha > 0$ rather than the half-Cauchy density $p(\alpha) = 2/\pi(\alpha^2 + 1), \alpha > 0$.[7] In addition, we will use modified Jeffreys' prior for μ and σ^2 under normality. That is, assuming that α, μ and σ^2 are independent a priori,

$$\pi(\alpha, \mu, \sigma^2) \propto \frac{1}{(\alpha + 1)^2} \frac{1}{\sigma^2}, \quad -\infty < \mu < \infty, \ \ \alpha, \sigma^2 > 0, \tag{3}$$

where consistent with standard use, we take σ^2 rather than $\sigma^{3/2}$. Thus, Equations (2) and (3) constitute the full DP model. The prior in Equation (3) is proper in α and improper in μ and σ^2, and a priori α, μ and σ^2 are independent.

Therefore, letting $g(y_i \mid \mu, \sigma^2)$, $i = 1, \ldots, n$, denote the normal density with mean μ and variance σ^2 and using Bayes' theorem in Equations (2) and (3), the joint posterior density is

$$\pi(\mu, \sigma^2, \alpha \mid \underset{\sim}{y}) \propto \frac{1}{\sigma^2(\alpha + 1)^2} g(y_1 \mid \mu, \sigma^2) \prod_{i=2}^{n} \left[\frac{1}{\alpha + i - 1} \left\{ \sum_{j=1}^{i-1} \delta_{y_j}(y_i) + \alpha g(y_i \mid \mu, \sigma^2) \right\} \right], \tag{4}$$

$-\infty < \mu < \infty, \sigma^2, \alpha > 0$. Let k denote the number of distinct values among $y_1, \ldots, y_n$. We write the posterior density in Equation (4) in a more convenient form in order to prove its propriety in Theorem 1. Let t_j, $j = 1, \ldots, k - 1$, denote the positions where the distinct values occur after the

first one in Equation (4) and let $\mathcal{T} = \{t_1, \ldots, t_{k-1}\}$. Then, letting $\bar{\delta}(y_i) = \sum_{j=1}^{i-1} \delta_{y_j}(y_i)/(i-1)$,

$$\pi(\mu, \sigma^2, \alpha \mid \underset{\sim}{y}) \propto \frac{1}{\sigma^2(\alpha+1)^2} g(y_1 \mid \mu, \sigma^2) \prod_{i=2}^{n} \left[\frac{1}{\alpha+i-1}\right] \times \left[\prod_{i \notin \mathcal{T}} \{(i-1)\bar{\delta}(y_i) + \alpha g(y_i \mid \mu, \sigma^2)\}\right] \left[\prod_{i \in \mathcal{T}} \alpha g(y_i \mid \mu, \sigma^2)\right], \quad -\infty < \mu < \infty, \ \sigma^2, \alpha > 0. \tag{5}$$

Note that once the y_i are observed, k and $\mathcal{T}$ are also observed.

Theorem 1: *If $k \geq 2$, the joint posterior density in Equation (5) $\pi(\mu, \sigma^2, \alpha \mid \underset{\sim}{y})$ is proper.*

Proof: See Appendix 2. The condition $k \geq 2$ in Theorem 1 is essentially minor. ■

Next, in order to obtain the exact prediction interval, we show how to obtain samples from the joint posterior density of $\underset{\sim}{y}_{ns}, \alpha, \mu, \sigma^2$ given $\underset{\sim}{y}_s$. We have

$$p(\underset{\sim}{y}_{ns}, \alpha, \mu, \sigma^2 \mid \underset{\sim}{y}_s) = p(\underset{\sim}{y}_{ns} \mid \alpha, \mu, \sigma^2, \underset{\sim}{y}_s) \pi(\alpha, \mu, \sigma^2 \mid \underset{\sim}{y}_s).$$

Once samples are taken from $\pi(\alpha, \mu, \sigma^2 \mid \underset{\sim}{y}_s)$, using the composition rule, samples are obtained from $p(\underset{\sim}{y}_{ns} \mid \alpha, \mu, \sigma^2, \underset{\sim}{y}_s)$. Thus, samples can be drawn from $p(\underset{\sim}{y}_{ns}, \alpha, \mu, \sigma^2 \mid \underset{\sim}{y}_s)$. However, as pointed out in the introductory remarks, if N is large this process is computationally prohibitive.

Letting k denote the number of distinct values in the observed data, Antoniak [8] showed that $p(k \mid \alpha) = s_n(k)\alpha^k \Gamma(\alpha)/\Gamma(\alpha+n), \alpha > 0$, where the $s_n(k)$, the absolute values of the Stirling numbers of the first kind,[9] are independent of μ and σ^2. Then, the joint posterior density of μ and σ^2 comes from the baseline model conditional on only the distinct values. That is, letting $y_1^*, \ldots, y_k^*$ denote the k distinct sample values ($k \geq 2$), we have

$$y_1^*, \ldots, y_k^* \mid k, \mu, \sigma^2 \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$$

with the prior in (3). Then, letting $\bar{y}_* = \sum_{i=1}^{k} y_i^*/k$ and $s_*^2 = \sum_{i=1}^{k}(y_i^* - \bar{y}_*)^2/(k-1)$, we have $\mu \mid \sigma^2, k, \bar{y}_*, s_*^2 \sim \text{Normal}(\bar{y}_*, \sigma^2/k)$ and $\sigma^{-2} \mid s_*^2, k \sim \text{Gamma}\{(k-1)/2, (k-1)s_*^2/2\}$. That is, $\sqrt{k}(\mu - \bar{y}_*)/s_* \mid \bar{y}_*, s_*^2, k \sim t_{k-1}$. Thus, it is trivial to draw μ and σ^2. It is not really trivial to draw α without using a special kind of prior; see [10] for a discussion of the gamma prior which was introduced earlier by Escobar and West.[11]

We present an improved method to draw α from its posterior density,

$$\pi(\alpha \mid k) \propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha+n)(\alpha+1)^2}, \quad \alpha > 0. \tag{6}$$

Transforming α according to $\rho = 1/(\alpha+1)$ (correlation in the DP) and simplifying Equation (6), we get

$$\pi(\rho \mid k) \propto \frac{(1-\rho)^{k-1}\rho^{n-k}}{\prod_{j=1}^{n-1}\{1-\rho+\rho j\}}, \quad 0 \leq \rho \leq 1. \tag{7}$$

Note that $\lim_{\rho \to 0} \pi(\rho \mid k) = 0 = \lim_{\rho \to 1} \pi(\rho \mid k)$, and $\pi(\rho \mid k)$ is well defined and differentiable everywhere in the closed interval $[0, 1]$.

Because the posterior density of ρ is not in a simple form, we use a one-dimensional grid method to draw samples from it, thereby avoiding Markov chain Monte Carlo methods (e.g. Metropolis sampler). The unit interval is simply divided into 100 sub-intervals of equal width, and the joint posterior density is approximated by a discrete distribution with probabilities proportional to the heights of the continuous distribution at the mid-points of these sub-intervals. Now, it is easy to draw a sample from this univariate discrete distribution of $\pi(\rho \mid k)$. It is efficient to remove sub-intervals with small probabilities (smaller than 10^{-6}); we call the others probable sub-intervals. To draw a single deviate, we first draw one of the probable sub-intervals. After we have obtained this sub-interval, a uniform random variable is drawn within this sub-interval. This is a standard jittering procedure which provides different deviates with probability one. This procedure works very well here because $\pi(\rho \mid k)$ is defined everywhere in the closed interval $[0, 1]$ and we have used it in several of our papers,[12, 13] to obtain samples from $\pi(\rho \mid k)$ and therefore $\pi(\alpha \mid k)$. The entire procedure is very fast as it takes just a few seconds to draw 10,000 values of α, μ, σ^2.

In theory it is easy to draw $\bar{y}_{ns}$. To each of the 10,000 iterates, simply fill in the values $y_{n+1}, \ldots, y_N$ (data augmentation). Using Equation (A2) of the generalized Polya urn scheme in Appendix 1, we have $y_{n+1} \mid \{\alpha, \mu, \sigma^2, \underset{\sim}{y}_s\} \sim (n/(\alpha+n))\bar{F}_n(y) + (\alpha/(\alpha+n))H$, and

$$y_{n+k+1} \mid \{\alpha, \mu, \sigma^2, \underset{\sim}{y}_s, y_{n+1}, \ldots, y_{n+k}\} \sim \frac{n+k}{\alpha+n+k}\bar{F}_{n+k}(y) + \frac{\alpha}{\alpha+n+k}H, \tag{8}$$

$k = 1, \ldots, N-n-1$, where $\bar{F}_{n+k}(y) = \sum_{i=1}^{n+k} F_{y_i}(y)/(n+k)$ and $\bar{F}_n(y)$ has $k = 0$. It is now easy to draw the nonsampled values one by one using Equation (8). The speed of this process is increased by drawing from $\bar{F}_{n+k}(y)$ using the multinomial distribution because there are repeats among the values already drawn.

Thus, we get 10,000 values of $\bar{Y}$; order these values and pick the 95% prediction interval to be $(\bar{y}_{(250)}, \bar{y}_{(9750)})$, where the values are arranged in increasing order. We call this interval the full (exact) Bayesian interval (FBI) and the method the full Bayesian method (FBM). Clearly, this procedure can be used for inference about quantiles. For each draw of the entire population compute the required quantile (e.g. median, Q) and then a 95% credible interval is $(Q_{(250)}, Q_{(9750)})$. However, when N is much larger than n, this procedure is computationally prohibitive as we will show in the examples. Typically n is much smaller than N and, therefore, the time to fit the model is negligible compare to the time to draw the $N-n$ nonsampled values from the DP model.

2.2. Approximate Bayesian prediction interval

We describe the approximate Bayesian predictive interval. Here we assume that the hyperparameters are fixed. Next, we present Theorem 2, which gives the required moments of $\bar{Y}$ conditional on the fixed hyperparameters, and assuming normality, we can obtain samples of the finite population mean, $\bar{Y}$, where we have used the samples of α, μ and σ^2 already drawn from the full DP model.

For fixed α, μ and σ^2, Binder [2] presented the PM and variance of $\bar{Y}$. We present more easily interpreted forms of the PM and variance of $\bar{Y}$. Let

$$\lambda = n(\alpha+N)/N(\alpha+n) \quad \text{and} \quad \phi = 1/(\alpha+n+1),$$

where $0 \le \lambda \le 1$ is a shrinkage parameter and ϕ is the posterior correlation. Momentarily, let $E'(\bar{Y}) = E(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)$ and $\mathrm{Var}'(\bar{Y}) = \mathrm{Var}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)$. Binder [2] stated the PM and variance of $\bar{Y}$ (Binder's formulae) and because they are needed later we state them in Theorem 2.

Theorem 2: *Assuming that the DP model holds,*

$$E'(\bar{Y}) = \lambda\bar{y} + (1-\lambda)\mu,$$
$$\mathrm{Var}'(\bar{Y}) = \lambda\left[(n-1)\phi(1-f)\frac{s^2}{n} + (1-\lambda)\left\{\phi(\bar{y}-\mu)^2 + (1-\phi)\frac{\sigma^2}{n}\right\}\right].$$

It is pertinent to discuss the behaviour of $E'(\bar{Y})$ and $\mathrm{Var}'(\bar{Y})$, stated in Theorem 2, for various choices of α. This is not discussed in [2] nor [3]. As α goes to zero, ϕ goes to $(n+1)^{-1}$ and λ goes to 1. In this case $E'(\bar{Y})$ becomes $\bar{y}$, the design-based estimator of $\bar{Y}$. Also, $\mathrm{Var}'(\bar{Y})$ becomes $\{(n-1)/(n+1)\}(1-f)s^2/n$ and if n is large this becomes the design-based estimator of the variance. Thus, we can retrieve the design-based prediction interval approximately. On the other hand, as α goes to infinity, ϕ goes to zero and λ goes to f, the sampling fraction. In this case $E'(\bar{Y})$ goes to $f\bar{y} + (1-f)\mu$, the prior prediction mean under normality and $\mathrm{Var}'(\bar{Y})$ becomes $(1-f)(\sigma^2/N)$, again the prior prediction variance under normality. Therefore, when α is large, draws are made mostly from the normal distribution and when α is small, draws are made mostly from the Polya posterior.[14, 15]

Therefore, it is easy to describe the ABI. As $y_{n+1}, \ldots, y_N \mid \underset{\sim}{y}_s$ are exchangeable, using Theorem 2,

$$\frac{\bar{Y} - E(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)}{\sqrt{\mathrm{Var}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)}} \sim \mathrm{Normal}(0,1) \tag{9}$$

asymptotically (as n and N go to infinity with $n < N$). We have used a result on central limit theorems for interchangeable (exchangeable) processes discussed by Blum et al.[16] It states that if $x_1, \ldots, x_N$ are exchangeable with mean μ, variance σ^2 (finite) and correlation ρ, then $\bar{X}$ is asymptotically normal with mean μ and variance $\rho\sigma^2$. [The actual variance is $(\sigma^2/N)(1 + (N-1)\rho)$ for any finite N.] In our case this is a very reasonable approximation for finite population sampling because N is generally large enough.

Therefore,

$$\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s \sim \mathrm{aNormal}\{E(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s),\ \mathrm{Var}(\bar{Y} \mid \mu, \sigma^2, \alpha, \underset{\sim}{y}_s)\}, \tag{10}$$

where 'aNormal' means asymptotically normal (as n and N go to infinity with $n < N$). With this normal approximation, we can proceed in the same manner as we did for the full Bayesian method; the difference is that we do not have to draw the nonsampled values.

We will make 10,000 draws from the posterior density of $\mu, \sigma^2, \alpha \mid \underset{\sim}{y}_s$ as described for the full Bayesian prediction interval. Then, for each draw we perform a data augmentation to obtain $\bar{Y}$ from the normal approximation in Equation (10). Thus, we get 10,000 values of $\bar{Y}$; order these values and pick the 95% prediction interval to be $(\bar{y}_{(250)}, \bar{y}_{(9750)})$, where the values are arranged in increasing order. We call this interval the ABI and the method the approximate Bayesian method (ABM). This is an enormous saving over the full Bayesian prediction interval because as we will see this approximation is very good for large population sizes where there are large computational savings. However, if quantiles are needed, the ABI must be abandoned and the exact method must be used. This is currently under investigation.

2.3. Empirical Bayes and exact moment prediction intervals

Like the design-based prediction intervals, we construct two additional approximate prediction intervals which are based on the DP. Starting with Equation (9) we simply use the formulae for the mean and variance in Theorem 2. The first method is empirical Bayes and the second obtains the exact mean and variance via numerical integration (not a sampling-based method). Theorem 3 gives the exact moments. For the empirical Bayes method we obtain the posterior modes of α, μ and σ^2. Then,

to obtain the prediction intervals, we assume normality. We will use the joint posterior density of μ, σ^2, α given $\underset{\sim}{y}_s$ in the exact and ABM.

First, we describe the empirical Bayes method. We will substitute posterior modes of μ, σ^2 and α into Equation (9). The posterior modes of μ and σ^2 are in closed forms and they are, respectively, $\hat{\mu} = \bar{y}_*$ and $\hat{\sigma}^2 = (k-1)s_*^2/(k+1), k > 1$. However, the posterior mode of α is a bit more involved. We study two procedures for finding the posterior mode of α, one is an iterative procedure and the other uses stochastic optimization.

Now, we describe the iterative procedure. Letting $\alpha = e^{\psi}$, the posterior density for ψ is

$$\pi(\psi \mid k) \propto \frac{e^{k\psi}}{(1+e^{\psi})^2 \prod_{j=1}^{n-1}(j+e^{\psi})}, \quad -\infty < \psi < \infty.$$

We note that $\pi(\psi \mid k)$ is logconcave (i.e. strongly unimodal with a unique mode). Then, taking the first derivative of $\pi(\psi \mid k)$ and setting it equal zero, we get the fixed-point solution

$$\psi = \ln\left\{\frac{k}{\sum_{j=1}^{n-1}(j+e^{\psi})^{-1} + 2(1+e^{\psi})^{-1}}\right\}.$$

Thus, starting at $\psi = 0$ after a few iterations we get the posterior mode $\hat{\psi}$ and therefore the posterior mode $\hat{\alpha} = e^{\hat{\psi}}$. This is similar to a procedure described in [17] which we have discovered independently.

The stochastic optimization to get the posterior mode is easy to perform. We have already shown how to get 10,000 iterates from the posterior density of ρ in Equation (7). Note that $\pi(\rho \mid k)$ is unimodal but not logconcave because it is the density of $\log(\rho)$ which is logconcave. Simply compute the $\pi(\rho \mid k)$ at each of the iterates, and take the value $\hat{\rho}$ where $\pi(\rho \mid k)$ has the largest value. Then $\hat{\alpha} = \hat{\rho}/(1-\hat{\rho})$ gives the posterior mode. Both the iterative procedure and the stochastic optimization give essentially the same posterior mode. We will call this interval the empirical Bayes interval (EBI) and the method to construct it the empirical Bayes method (EBM).

Second, we describe the integration to obtain the exact moments (mean and variance). In Theorem 3 we obtain almost the complete forms of the moments.

Theorem 3: *Assuming that the DP model holds and* $k \geq 4$,

$$E(\bar{Y} \mid \underset{\sim}{y}_s, k) = E(\lambda \mid k)\bar{y} + \{1 - E(\lambda \mid k)\}\bar{y}_* \quad \textit{and} \quad \mathrm{Var}(\bar{Y} \mid \underset{\sim}{y}_s, k) = V_1 + V_2,$$

$$V_1 = (n-1)(1-f)\frac{s^2}{n}E(\lambda\phi \mid k) + \left\{(\bar{y}-\bar{y}_*)^2 + \frac{(k-2)s_*^2}{k(k-3)}\right\}E\{\lambda(1-\lambda)\phi \mid k\}$$
$$+ \frac{(k-1)s_*^2}{(k-3)n}E\{\lambda(1-\lambda)(1-\phi) \mid k\},$$

$$V_2 = (\bar{y}-\bar{y}_*)^2 \mathrm{var}(\lambda \mid k) + \frac{(k-2)s_*^2}{k(k-3)}E\{(1-\lambda)^2 \mid k\},$$

where expectations are taken over the posterior density of α.

Proof: See Appendix 3. The condition $k \geq 4$ in Theorem 3 is essentially minor. ■

Finally, the integration over α can be obtained either by numerical or Monte Carlo techniques. We use the latter with the 10,000 draws we already made from the posterior density of α, described in

Equation (6), which we write fully as

$$\pi(\alpha \mid k) = \frac{\alpha^{k-1}\left\{\prod_{j=1}^{n-1}(j+\alpha)\right\}^{-1}(1+\alpha)^{-2}}{\int_0^\infty \alpha^{k-1}\left\{\prod_{j=1}^{n-1}(j+\alpha)\right\}^{-1}(1+\alpha)^{-2}\,\mathrm{d}\alpha}, \quad \alpha > 0.$$

Letting $g(\alpha)$ be any integrable function of α,

$$E\{g(\alpha) \mid k\} = \frac{\int_0^\infty g(\alpha)\alpha^{k-1}\left\{\prod_{j=1}^{n-1}(j+\alpha)\right\}^{-1}(1+\alpha)^{-2}\,\mathrm{d}\alpha}{\int_0^\infty \alpha^{k-1}\left\{\prod_{j=1}^{n-1}(j+\alpha)\right\}^{-1}(1+\alpha)^{-2}\,\mathrm{d}\alpha}.$$

Then, a good Monte Carlo estimate of $E\{g(\alpha) \mid k\}$ is

$$\hat{E}\{g(\alpha) \mid k\} = \sum_{h=1}^{10000} w_h g(\alpha_h),$$

where $w_h \propto \alpha_h^{k-1}\{\prod_{j=1}^{n-1}(j+\alpha_h)\}^{-1}(1+\alpha_h)^{-2}, h = 1, \ldots, 10,000$, and $\alpha_h \overset{\text{iid}}{\sim} \pi(\alpha \mid k)$. We apply this method to each of the required integrals. The computation of the expectations took only a few seconds. We will call this interval the exact moment interval (EMI) and the method to construct it the exact moment method (EMM).

3. Examples and simulation study

To compare our five intervals/methods, we discuss 14 examples and a simulation study. We are particularly interested in the comparison between the approximate Bayesian method (ABI/ABM) and the full (exact) Bayesian method (FBI/FBM) but we also make comparisons with the other intervals/methods: design-based (DBI/DBM), empirical Bayes (EBI/EBM) and exact moment (EMI/EMM).

In the 14 examples the population sizes vary considerably. The first 13 examples are on the third National Health and Nutrition Examination Survey (NHANES III). These are the data on body mass index where we assume that equivalent simple random samples are taken from 13 states. We are particularly interested in females older than 45 years because one of our projects is on obesity of women who have gone beyond the onset of menopause. The population sizes for the obesity study are around one million and the sample sizes are considerably smaller making the prediction problem challenging in terms of time. The 14th dataset is taken from Aitkin [18] on income which he used to discuss finite population sampling. This is a much smaller population which creates little difficulty in terms of time for the full (exact) Bayesian method.

The examples show that the results from the approximate method are similar to those from the full (exact) Bayesian method. So we discuss this further in the simulation study. Because it takes a very long time to do the computations for the NHANES III examples, we have restricted our simulation study to population sizes similar to Aitkin [18] which is moderately large.

3.1. Examples

In Table 1 we present a comparison of four methods (DBM, EBM, EMM and ABM) by examples. We have used the PM and PSD of the finite population mean. There are some differences among the four methods. Sometimes the differences are large. It is surprising that EBM tends to have larger PMs but, as expected, the PSDs from EBM are smaller. For example, for the income data the PMs are about the same for DBM and EMM and the PSDs are close, the PM for EBM is a bit larger and the PSD is a bit

Table 1. Comparison of PM and posterior standard deviation (PSD) of the finite population mean for 14 examples by methods.

	DBM		EBM		EMM		ABM	
n; *N*	PM	PSD	PM	PSD	PM	PSD	PM	PSD
25; 608491	25.880	1.205	26.254	1.046	25.879	1.158	25.807	1.307
556; 4453263	28.045	0.272	28.185	0.276	28.046	0.271	28.131	0.275
162; 2704478	28.086	0.490	28.160	0.495	28.088	0.487	28.250	0.508
86; 1985501	28.860	0.676	29.113	0.657	28.862	0.668	29.038	0.701
47; 1086648	26.213	0.844	26.493	0.799	26.216	0.826	26.522	0.904
80; 1562869	28.150	0.642	28.297	0.632	28.152	0.634	28.339	0.676
59; 947239	27.458	0.669	27.558	0.667	27.460	0.659	27.628	0.725
322; 3310865	28.009	0.339	28.086	0.342	28.010	0.338	28.079	0.343
83; 1949322	27.229	0.687	27.451	0.663	27.230	0.678	27.382	0.708
129; 2358615	26.690	0.534	26.803	0.534	26.692	0.530	26.871	0.552
45; 190472	28.444	1.131	30.324	1.089	28.447	1.106	28.703	1.259
240; 2524603	28.521	0.361	28.574	0.364	28.522	0.360	28.602	0.369
64; 776246	27.031	0.683	27.619	0.663	27.035	0.672	27.247	0.711
48; 648	67.075	3.471	70.775	2.458	67.076	3.385	67.845	3.518

Note: PM is the posterior mean; PSD is the posterior standard deviation. The first 13 examples are from NHANES III and the 14th one is a data set on income.[18] DBM is the design-based method, EBM is the empirical Bayes method, EMM is the exact moment method and ABM is the approximate Bayesian method.

smaller. Except for the first example the PMs under EMM are larger than those under ABM and for all examples PSDs under EMM are smaller than those under ABM, but the differences are small. As expected, in all examples DBM, EBM and EMM have smaller PSDs than ABM.

In Table 2 we have first assessed normality of the posterior distribution of $\bar{Y}$ using the Kolmogorov–Smirnov test (KST). The one sample KST shows that there is no reason to dispute normality for all 14 examples. The smallest p-value is .158 for ABM and .280 for FBM. We have also used the two-sample KST to compare the posterior distributions of $\bar{Y}$ under ABM and FBM. For the 14 examples the p-values are .985, .979, .742, .342, .268, .742, .621, .849, .672, .865, .865, .560, .605, .757. Side-by-side box plots (omitted) of the posterior densities under the ABM and FBM are symmetric and they look very similar except ABM tends to have slightly more outliers. Thus, there is no reason to believe that the posterior distributions of $\bar{Y}$ under ABM and FBM are different, and in fact, they have reasonably approximate normal distributions. This is true for all 14 examples. In Table 2 we have also studied ABM and FBM. We have compared the PMs, PSDs and the 95% credible intervals of $\bar{Y}$. Except for Examples 9, 11 and 13 PSDs under ABM are smaller than those under FBM. Otherwise, the 95% credible intervals are very close. Although not particularly relevant, one can see that, except for Examples 1, 5, 11 and 14 (much smaller sample size), the PSDs are larger (similar for both methods) than for the other examples. Therefore, ABM is a good competitor to FBM which needs much more computational effort.

In Table 3 we have compared the time (hours) it takes to do the computation on our Linux Computational Node with 2.70 GHz and 8 CPU Cores. For the ABM the real time for the computation of all 14 examples combined is just 8.8 seconds. This time includes the computations of the design method, the empirical Bayes method and the method based on the exact moments. These are not included in the computations for the exact method. The computation to obtain the samples from the joint posterior density of μ, σ^2, α is common to both methods. The Kolmogorov–Smirnov tests for normality are included in both procedures. There is enormous variation in the times for the various examples. For Example 2 ($N = 4,453,263$) the time for the computation is 44.311 h and in Example 11 ($N = 190,472$) the time is 1.895 h. Example 14 ($N = 648$) took just 21.6 seconds (or 0.006 hour). For the income data, Aitkin [18] reported the DBI as (60.6, 73.6) and a Bayesian bootstrap interval based on 10, 000 bootstrap samples as (60.6, 74.2) which are slightly narrower than the ones we have in Table 2 but substantially overlapping on the left. The ABI and FBI are, respectively, (61.1, 74.9) and (61.2, 75.2).

Table 2. Comparison of the ABM and the full (exact) Bayesian method (FBM) for posterior inference of the finite population mean for 14 examples.

	ABM				FBM			
$n; N^a$	PM	PSD	95% CI	Pval	PM	PSD	95% CI	Pval
25; 0.6	25.807	1.307	(23.317, 28.484)	.158	25.794	1.319	(23.233, 28.226)	.881
556; 4.5	28.131	0.275	(27.588, 28.658)	.982	28.155	0.279	(27.587, 28.665)	.996
162; 2.7	28.250	0.508	(27.296, 29.296)	.604	28.259	0.522	(27.239, 29.207)	.429
86; 2.0	29.038	0.701	(27.670, 30.392)	.731	29.056	0.708	(27.817, 30.472)	.450
47; 1.1	26.522	0.904	(24.792, 28.349)	.725	26.578	0.928	(24.842, 28.427)	.280
80; 1.6	28.339	0.676	(27.020, 29.676)	.512	28.358	0.713	(26.987, 29.714)	.491
59; 0.9	27.628	0.725	(26.302, 29.125)	.984	27.640	0.742	(26.231, 29.098)	.689
322; 3.3	28.079	0.343	(27.416, 28.753)	.930	28.070	0.361	(27.401, 28.820)	.818
83; 1.9	27.382	0.708	(25.967, 28.748)	.985	27.377	0.688	(26.014, 28.657)	.632
129; 2.4	26.871	0.552	(25.799, 27.962)	.956	26.873	0.561	(25.877, 27.967)	.903
45; 0.2	28.703	1.259	(26.198, 31.136)	.316	28.656	1.214	(26.087, 30.985)	.720
240; 2.5	28.602	0.369	(27.875, 29.323)	.984	28.606	0.372	(27.866, 29.294)	.660
64; 0.8	27.247	0.711	(25.827, 28.634)	.543	27.227	0.688	(25.930, 28.535)	.464
40; 644	67.845	3.518	(61.051, 74.903)	.410	67.868	3.558	(61.173, 75.169)	.763

Note: PM is the posterior mean; PSD is the posterior standard deviation; CI is the credible interval; Pval refers to the Kolmogorov test for normality. [a] Except for the last example N must be multiplied by 10^6; see the note to Table 1 for the exact population sizes. The procedure uses 10,000 draws from the approximate posterior density. The BMI data set has a single US state for females older than 45 years from NHANES III and the last example is on the income data.[18]

Table 3. Comparison of the times (hours) for the ABM and the full (exact) Bayesian method (FBM) to perform the computations for the finite population mean by example.

$n; N$	FBM
25; 608491	6.055
556; 4453263	44.311
162; 2704478	26.910
86; 1985501	19.756
47; 1086648	10.812
80; 1562869	15.551
59; 947239	9.425
322; 3310865	32.944
83; 1949322	19.396
129; 2358615	23.469
45; 190472	1.895
240; 2524603	25.120
64; 776246	7.724
48; 648	0.006

Note: The total time it took to compute all 14 examples just 8.8 s using the ABM. The computations to obtain the samples from the joint posterior density of μ, σ^2, α is common to both methods. The first 13 examples are from NHANES III and the 14th one is a data set on income.[18]

Thus, for population sizes of 1000 the time to run EBM is not significant. However, the time to run population sizes of $1,000,000$ is intolerable, and therefore, an approximation such as the one we have developed is useful. More importantly the posterior distributions of the finite population mean under ABM and FBM are approximate normal distributions and posterior inferences are similar. So it is reasonable to use ABI for large populations and the FBI for small populations. Thus, using a simulation study we have investigated the performance of the ABI and the FBI for moderate size populations.

Table 4. Simulation Study: Comparison of the ABM and exact Bayesian method (FBM) using five-number summary of ratios of PM, PSD and end points of 95% prediction intervals by sample size and population size

n, N	Ratio	Min	Q1	Med	Q3	Max	AVG	STD
10, 100	PM	−.005	−.001	.0001	.0006	.0122	.0001	0.0011
	PSD	−.039	−.009	.0006	.0092	.0523	.0003	0.0133
	Bot	−.056	−.003	.0011	.0053	.1076	.0011	0.0093
	Top	−.035	−.003	.0006	.0042	.0224	.0007	0.0057
25, 250	PM	−.003	−.000	.0000	.0004	.0020	.0000	0.0006
	PSD	−.041	−.008	−.000	.0077	.0380	−.000	0.0110
	Bot	−.020	−.002	.0005	.0032	.0274	.0005	0.0041
	Top	−.017	−.002	.0004	.0025	.0161	.0004	0.0034
50, 500	PM	−.002	−.000	−.000	.0004	.0022	−.000	0.0006
	PSD	−.029	−.006	.0008	.0078	.0309	.0007	0.0106
	Bot	−.012	−.003	−.001	.0016	.0116	−.001	0.0036
	Top	−.009	−.003	−.001	.0015	.0113	−.001	0.0031
100, 1000	PM	−.001	−.000	.0000	.0003	.0013	.0000	0.0004
	PSD	−.033	−.006	.0006	.0068	.0261	.0005	0.0095
	Bot	−.009	−.002	−.001	.0012	.0083	−.000	0.0025
	Top	−.007	−.002	−.000	.0010	.0086	−.000	0.0022

Note: There are 1000 runs in the simulations and the PM, PSD and the 95% prediction interval is computed for each run.

3.2. Simulation study

To study the small population properties of the ABI, we have performed a simulation study. We have drawn our sample using a Parzen–Rosenblatt kernel density estimate with a window width obtained using the income data. Let $y_1, \ldots, y_n$ denote the sample of size $n = 40$ observations from the income data. So that

$$\widehat{f(y)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_o}\phi\left(\frac{y - y_i}{h_o}\right), \quad -\infty < y < \infty,$$

where h_o is the optimal window width [19] and $\phi(\cdot)$ is the standard normal density. We have drawn a random sample (population) of size N from this kernel density and we have taken a simple random sample of size n from this selected population. We have considered sample sizes and population sizes $(n, N) = (10, 100), (25, 250), (50, 500), (100, 1000)$. We have drawn 1000 samples at each of the four design points (N, n), and we have run the computation in exactly the same manner as we have described for all five methods (DBM, EBM, EMM, ABM and FBM). Thus, we know the finite population mean.

In Table 4 we have a comparison of ABM and FBM using relative difference between their PMs, PSDs and end points of 95% credible intervals of the finite population mean of the 1000 simulation runs. For example, for the PMs we compute 1000 ratios [i.e. Ratio $= (PM_{\text{abm}} - \text{PM}_{\text{fbm}})/\text{PM}_{\text{fbm}}$] and we have presented the five-number summary as well as their average (AVG) and standard deviation (STD). For all values of (n, N) these quantities are reasonably small [except perhaps at $(n, N) = (10, 100)$ where max $= .1076$] suggesting that these two methods are very close except perhaps at $n = 10$, $N = 100$.

In Table 5 we perform the Kolmogorov–Smirnov test (KST) of equality of distribution (ED) of the two posterior distributions and the tests of normality for these posterior densities (AN for ABM and EN for FBM). We have presented the five number summary of the p-values as well as the fifth percentile ($P5$). Clearly, the test of normality fails at ($n = 10$, $N = 100$) (i.e. the posterior distribution of $\bar{Y}$ is not normal) and the equality of the posterior distributions is slightly questionable. It seems reasonable to conclude that about $n = 50$ $N = 500$ or greater, there is normality for both methods and they have the same distributions (the p-values are larger than .05 for $Q1$). For smaller values of (n, N) one needs to use the exact method; this runs very fast anyway.

Table 5. Simulation Study: Five-number summary of the *p*-values over the 1000 runs by sample size (n), population size (N) and test.

n, *N*	Test	P5	Min	Q1	Med	Q3	Max
10, 100	ED	.0001	.0926	.3646	.6318	.8519	.9996
	AN	.0000	.0000	.0000	.0002	.0008	.0221
	EN	.0000	.0000	.0000	.0001	.0003	.0307
25, 250	ED	.0011	.0926	.3646	.6258	.8615	.9999
	AN	.0019	.0463	.1694	.3053	.4970	.9946
	EN	.0000	.0010	.0253	.0968	.2243	.9622
50, 500	ED	.0018	.0645	.2284	.4708	.7326	.9996
	AN	.0547	.2743	.5364	.7608	.9062	1.000
	EN	.0000	.0073	.0581	.1926	.4254	.9989
100, 1000	ED	.0016	.0862	.2564	.4762	.7092	.9975
	AN	.1404	.3615	.6669	.8517	.9478	.9999
	EN	.0006	.0205	.1001	.2428	.4565	.9958

Note: There are 1000 runs in the simulations and for each run the Kolmogorov–Smirnov test is performed of each of 10000 values in the sampling-based method. For each simulation run, three tests are performed (ED: equality of the two distributions; AN: equality of approximation and the normal distributions; EN: equality of the exact and normal distributions). We have included the fifth percentile (P5).

Table 6. Simulation Study: Comparison of coverage (C), Rbias, PRMSE and Wid of the five prediction intervals (DBI, EBI, EMI, ABI, FBI) by sample size (*n*), population size (*N*).

n, *N*	Method	Rbias	PRMSE	*C*	Wid
10, 100	DBI	0.018; 0.003	4.339; 0.095	0.833; 0.012	10.74; 0.191
	EBI	0.015; 0.004	3.987; 0.110	0.508; 0.016	5.805; 0.116
	EMI	0.018; 0.003	4.157; 0.094	0.795; 0.013	9.715; 0.172
	ABI	0.016; 0.003	4.524; 0.100	0.854; 0.011	11.54; 0.211
	FBI	0.016; 0.003	4.524; 0.100	0.851; 0.011	11.54; 0.210
25, 250	DBI	−.010; 0.002	3.292; 0.041	0.907; 0.009	8.874; 0.056
	EBI	−.015; 0.002	3.193; 0.053	0.750; 0.014	6.974; 0.047
	EMI	−.010; 0.002	3.225; 0.041	0.892; 0.010	8.522; 0.054
	ABI	−.016; 0.002	3.383; 0.045	0.902; 0.009	8.992; 0.060
	FBI	−.016; 0.002	3.384; 0.045	0.901; 0.009	8.998; 0.061
50, 500	DBI	−.001; 0.001	4.104; 0.039	0.933; 0.008	11.70; 0.036
	EBI	0.004; 0.001	3.802; 0.040	0.922; 0.008	10.42; 0.028
	EMI	−.000; 0.001	4.055; 0.040	0.927; 0.008	11.46; 0.035
	ABI	0.003; 0.001	3.924; 0.035	0.953; 0.007	11.53; 0.030
	FBI	0.003; 0.001	3.925; 0.035	0.951; 0.007	11.53; 0.030
100, 1000	DBI	0.002; 0.001	2.871; 0.027	0.954; 0.007	8.330; 0.018
	EBI	0.004; 0.001	2.743; 0.027	0.955; 0.007	7.919; 0.015
	EMI	0.002; 0.001	2.851; 0.028	0.952; 0.007	8.236; 0.018
	ABI	0.004; 0.001	2.766; 0.025	0.960; 0.006	8.191; 0.015
	FBI	0.004; 0.001	2.766; 0.025	0.963; 0.006	8.188; 0.015

Notes : There are 1000 runs in the simulations and for each run the 95% credible intervals, Bias, posterior root-mean-squared error and credible incidence (whether an interval contains the true value) and wid of the credible intervals are calculated. The first number in each entry is the average over the 1000 runs and of the second number is the standard error of the average. The first three methods are based only on the mean and standard deviation with an assumption of normality. The last two methods are based on Monte Carlo with an approximation of normality for the first of these two.

In Table 6 we investigate consistency properties such as relative bias, mean-squared error, width of 95% credible intervals and the coverage of the intervals. The relative bias is $\text{Rbias} = (\text{PM} - \bar{Y})/\bar{Y}$, the posterior root-mean-squared error is $\text{PRMSE} = \sqrt{(\text{PM} - \bar{Y})^2 + \text{PSD}^2}$, the width (wid) is the difference between the upper end and the lower end of the 95% credible interval, and the credible incidence is 1 if the 95% credible interval contains the true value and 0 otherwise. We have done this for all five prediction intervals (DBI, EBI, EMI, ABI and FBI). We have taken the average of these quantities over the 1000 runs. The average of the credible incidences is the coverage (*C*). The

95% credible intervals for DBM, EBM and EMM are obtained as $PM \pm 1.96PSD$. For all methods the relative bias is negligible especially for larger (n, N) and the PRMSE also gets smaller as (n, N) increases. The methods are reasonably similar, but as expected, EBM has smaller PRMSE and the 95% credible intervals are too short. The coverage provided by the EBI is too small for values of (n, N) smaller than $(100, 1000)$. DBI also has coverage smaller than the nominal value of 95%. The ABI and the FBI have similar coverages for all values of (n, N). However, as (n, N) increases the coverages for all methods approach the nominal value. It is very interesting that ABI and FBI are very close over all measures.

4. Concluding remarks

In this paper our goal has been to obtain a Bayesian prediction interval for the finite population mean when the population size is much larger than the sample size under the DP. When the full Bayesian method is used, the computation is prohibitive. So we have obtained an approximate Bayesian prediction interval which is virtually the same as the full (exact) Bayesian prediction interval for relatively large population with substantially reduced computational time. We have also made comparisons with some standard methods (design-based, empirical Bayes and exact moments with an assumption of normality).

Parametric assumptions can be tenuous in many applications (e.g. survey sampling) because there are typically gaps and ties in such data. Generally, such gaps and ties are not taken into consideration seriously. Thus, a nonparametric procedure is desirable especially when a prediction interval is needed for a population mean or quantile. Under the DP using asymptotic theory, Binder [2] and Lo [3] obtained the standard design-based prediction interval under simple random sampling. Under the DP for the finite population values, the DBI does not take all features of the DP under consideration. We have shown how to obtain the exact Bayesian prediction interval. We note that this work has enormous potential for many complex surveys which are, in fact, naturally nonparametric.

We have used several numerical examples and a simulation study with simple random samples drawn from the Parzen-Rosenblatt kernel density estimator. We have one recommendation. The exact Bayesian method should be used when prediction is to be done for small to moderate populations (size less than 500) and the ABM should be used for much larger populations. The exact method must be used if quantile estimation is needed, but the computational time can be prohibitive for large populations. This latter situation needs further study.

Finally, as in the work by Binder [2] and Lo [3], it is easy to extend our work for stratification. It may be possible to considerably reduce the time for computation using the stick-breaking algorithm,[20] and we will report on this research elsewhere.

Acknowledgments

The authors are grateful to the two reviewers for their careful reading of the manuscript and their suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann Stat. 1973;1:209–230.
[2] Binder DA. Non-parametric Bayesian models for samples from finite populations. J R Statist Soc, Ser B. 1982;44:388–393.
[3] Lo AY. Bayesian statistical inference for sampling a finite population. Ann Stat. 1986;14:1226–1233.
[4] Lo AY. A Bayesian bootstrap for a finite population. Ann Stat. 1988;26:1684–1695.
[5] Hjort NL, Holmes C, Muller P, Walker SG. Bayesian nonparametrics. New York: Chapman and Hall; 2010.
[6] Nandram B, Yin J. Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline. Statist Methodol. 2016;28:1–17. doi:10.1016/j.stamet.2015.07.003.

[7] Polson NG, Scott JG. On the half-cauchy prior for a global scale parameter. Bayesian Anal. 2012;7:887–902.
[8] Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat. 1974;2:1152–1174.
[9] Abramowitz M, Stegun IA. Handbook of mathematical functions with formulas, graphs and mathematical tables. New York: Dover Publications; 1972.
[10] Nandram B, Choi JW. Nonparametric Bayesian analysis of a proportion for small area under nonignorable nonresponse. J Nonparametric Stat. 2004;16:821–839.
[11] Escobar MD, West M. Bayesian density estimation and inference using mixtures. J Amer Statist Assoc. 1995;90(430):577–588.
[12] Molina I, Nandram B, Rao JNK. Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. Ann Appl Stat. 2014;8(2):852–885.
[13] Nandram B, Toto MCS, Choi JW. A Bayesian benchmarking of the Scott-Smith model for small areas. J Statist Comput Simul. 2011;81(11):1593–1608.
[14] Ghosh M, Meeden G. Bayesian methods for finite population sampling. Cambridge: Cambridge University Press; 1997.
[15] Nelson D, Meeden G. Noninformative nonparametric quantile estimation for simple random samples. J Statist Plann Inference. 2006;136:53–67.
[16] Blum JR, Chernoff H, Rosenblatt M, Teicher H. Central limit theorems for interchangeable processes. Canad J Math. 1958;10:222–229.
[17] Liu JS. Nonparametric hierarchical Bayes via sequential imputations. Ann Stat. 1996;24:911–930.
[18] Aitkin M. Statistical inference: an integrated Bayesian/likelihood approach. New York: Chapman & Hall; 2010.
[19] Silverman BW. Density estimation for statistics and data analysis. New York: Chapman and Hall; 1986.
[20] Sethuraman J. A constructive definition of Dirichlet priors. Statist Sin. 1994;4:639–650.
[21] Blackwell D, MacQueen JB. Ferguson distributions via Polya Urn schemes. Ann Stat. 1973;1:353–355.
[22] Chaudhuri S, Ghosh M. Empirical likelihood for small area estimation. Biometrika. 2011;87:633–649.
[23] Ferguson TS. Bayesian density estimation by mixtures of normal distributions. Recent advances in statistics: papers in honor of Herman Chernoff on his sixtieth birthday. New York: Academic Press; 1983. p. 287–302.

Appendix 1. A brief review of the DP

We consider a finite population of N units with values $y_1, \ldots, y_N$. We describe key features of the DP,

$$y_1, \ldots, y_N \mid G \overset{\text{iid}}{\sim} G, G \mid H \sim \text{DP}(\alpha, H), \tag{A1}$$

where α is a concentration parameter, G is the unknown random probability measure, H is the baseline distribution (parametric) and is typically taken as the normal distribution, although other continuous distributions can be used. Ferguson [1] established the existence of the DP and showed that it is discrete with probability one. Here G is a random distribution which varies around H, a centre. So that the DP captures an uncountable number of models (hence the word nonparametric, somewhat inaccurate terminology). A key property of the DP is that $y_1, \ldots, y_N$ are exchangeable (i.e. the y_i have the same marginal distribution and they are equi-correlated).

Let μ_o and σ_o^2 denote the mean and variance of y_i under the baseline measure H. Momentarily, we assume that α, μ_o and σ_o^2 are fixed and so it is convenient to drop the conditioning on them. It is interesting that the DP is a generalized Polya urn scheme. Blackwell and MacQueen [21] showed how to obtain the joint distribution of $y_1, \ldots, y_N$ by integrating out G to get the generalized Polya urn scheme,

$$y_{k+1} \mid y_1, \ldots, y_k \sim \frac{k}{\alpha + k}\bar{F}(y) + \frac{\alpha}{\alpha + k}H, \quad k = 1, \ldots, N-1, \tag{A2}$$

where $y_1 \sim H$, $\bar{F}(y) = \frac{1}{k}\sum_{i=1}^{k} F_{y_i}(y)$, and $F_{y_i}(y)$ is the cdf of a point mass at y_i.

Ferguson [1] also obtained the posterior distribution of G which, letting $\bar{F}_o(y) = (1/n)\sum_{i=1}^{n} F_{y_i}(y)$, is

$$G \mid \underset{\sim}{y_s} \sim \text{DP}(\alpha^*, H^*), \tag{A3}$$

where $\alpha^* = \alpha + n$ and $H^* = (n/(\alpha + n))\bar{F}_o(y) + (\alpha/(\alpha + n))H$, a conjugate posterior density; see [5] for a heuristic demonstration.

For completeness we mention the Dirichlet process mixture (DPM) model which is

$$y_i \mid \mu_i \overset{\text{ind}}{\sim} f(y \mid \mu_i, \psi), \quad i = 1, \ldots, N,$$

$$\mu_1, \ldots, \mu_N \mid G \overset{\text{iid}}{\sim} G \quad \text{and} \quad G \mid H_{\underset{\sim}{\theta}} \sim \text{DP}\{\alpha, H_{\underset{\sim}{\theta}}\}.$$

This model is currently receiving a lot of attention and it has been used in countless applications; see, for example, Chaudhuri and Ghosh [22] for small area estimation and Nandram and Choi [10] for small area estimation with nonignorable nonresponse.

It is worth noting that in the DPM the parametric distribution, $f(y \mid \mu_i, \psi)$, has to be specified while no such specification is needed in the DP model. Besides, in practice, inference is likely to be sensitive to the specification of $f(y \mid \mu_i, \psi)$ and model diagnostics will be needed. Nevertheless, the whole idea is that the discreteness [1] of G in the DP is removed by using the DPM model.[23] This is advantageous for many applications (e.g. estimation of a density function), but with a simple random sample, the DPM appears to be an over kill. Moreover, it ensures that each observation comes from a different distribution and is not appropriate for a simple random sample without additional information (e.g. covariates). For data with gaps and ties, it seems more appropriate to use the DP, which is more nonparametric than the DPM model. Clearly, the DP model is still not fully nonparametric and for a practical solution a prior distribution must be assumed for α and $\underset{\sim}{\theta}$. There are no difficulties in finding a prediction interval under the DPM because the sampling process is parametric. However, if we replace the sampling process by a DP, our method for the DP can be used. This is under investigation.

Appendix 2. Proof of Theorem thm1

Using the form of the joint posterior density in Equation (5) and noting that $\alpha/(\alpha+i-1) \le 1,\ i=1,\ldots,n$, and $\bar{\delta}(y_i) = \sum_{j=1}^{i-1} \delta_{y_j}(y_i)/(i-1) \le 1,\ i=2,\ldots,n$, we have

$$\frac{1}{\sigma^2(\alpha+1)^2} g(y_1 \mid \mu,\sigma^2) \prod_{i=2}^{n} \left[\frac{1}{\alpha+i-1}\right] \left[\prod_{i\notin\mathcal{T}} \{(i-1)\bar{\delta}(y_i) + \alpha g(y_i \mid \mu,\sigma^2)\}\right] \left[\prod_{i\in\mathcal{T}} \alpha g(y_i \mid \mu,\sigma^2)\right]$$

$$\le \frac{1}{\sigma^2(\alpha+1)^2} g(y_1 \mid \mu,\sigma^2) \left[\prod_{i\notin\mathcal{T}} [\max\{\bar{\delta}(y_i), g(y_i \mid \mu,\sigma^2)\}]\right] \left[\prod_{i\in\mathcal{T}} g(y_i \mid \mu,\sigma^2)\right]$$

Without loss of generality, assume that $\max\{\bar{\delta}(y_i), g(y_i \mid \mu,\sigma^2)\} \le 1$. [It is also possible for $\max\{\bar{\delta}(y_i), g(y_i \mid \mu,\sigma^2)\} \le g(y_i \mid \mu,\sigma^2)$, but it does not matter.]

Therefore, we only need to show that

$$\int_0^\infty \int_{-\infty}^{\infty} \int_0^\infty \frac{1}{\sigma^2(\alpha+1)^2} g(y_1 \mid \mu,\sigma^2) \left[\prod_{i\in\mathcal{T}} g(y_i \mid \mu,\sigma^2)\right] d\alpha\, d\mu\, d\sigma^2 < \infty. \tag{A4}$$

Integrating out α (proper prior), we now only need to show that

$$\int_0^\infty \int_{-\infty}^{\infty} \frac{1}{\sigma^2} g(y_1 \mid \mu,\sigma^2) \left[\prod_{i\in\mathcal{T}} g(y_i \mid \mu,\sigma^2)\right] d\mu\, d\sigma^2 < \infty. \tag{A5}$$

Looking at Equation (A5), we only need to prove that $\pi_k(\mu,\sigma^2 \mid \underset{\sim}{y_k}) \propto (1/\sigma^2)[\prod_{i=1}^{k} g(y_i \mid \mu,\sigma^2),\ k \ge 2,\ -\infty < \mu < \infty,\ \sigma^2 > 0$, where $\underset{\sim}{y_k} = (y_1,\ldots,y_k)'$ is the vector of the k distinct values, is proper. But, letting $\bar{y}_k = \sum_{i=1}^{k} y_i/k$ and $s_k^2 = \sum_{i=1}^{k}(y_i-\bar{y}_k)^2/(k-1)$, for $k \ge 2$ it is well known that $(k-1)s_k^2/\sigma^2 \mid \underset{\sim}{y_k} \sim \chi^2_{k-1}$ and $\mu \mid \sigma^2, \underset{\sim}{y_k} \sim \text{Normal}(\bar{y}_k, \sigma^2/k)$. Thus, $\pi_k(\mu,\sigma^2 \mid \underset{\sim}{y_k})$ is proper and, therefore, $\pi_(\mu,\sigma^2,\alpha \mid \underset{\sim}{y})$ is proper.

Appendix 3. Proof of Theorem thm3

We integrate $\Omega = (\mu,\sigma^2,\alpha)$ out of the moments, stated in Theorem 2, using the conditional mean and variance formulas. That is,

$$E(\bar{Y} \mid \underset{\sim}{y_s}) = E\{E(\bar{Y} \mid \underset{\sim}{y_s}, \Omega)\}, \tag{A6}$$

$$\text{Var}(\bar{Y} \mid \underset{\sim}{y_s}) = V_1 + V_2, \quad V_1 = E\{\text{Var}(\bar{Y} \mid \underset{\sim}{y_s}, \Omega)\}, \quad V_2 = \text{Var}\{E(\bar{Y} \mid \underset{\sim}{y_s}, \Omega)\}, \tag{A7}$$

where $E(\bar{Y} \mid \underset{\sim}{y_s}, \Omega)$ and $\text{Var}(\bar{Y} \mid \underset{\sim}{y_s}, \Omega)$ are given by Theorem 2. We need to determine V_1 and V_2. It is worth noting that α and (μ,σ^2) are independent a posteriori with $(k-1)s_*^2/\sigma^2 \mid \underset{\sim}{y_*}, k \sim \chi^2_{k-1}$ and $\sqrt{k}(\mu-\bar{y}_*)/s_*^2 \mid \underset{\sim}{y_*}, k \sim t_{k-1}$, a Student's t density. Then, $\text{Var}(\mu \mid \underset{\sim}{y_s}, k) = (k-2)s_*^2/k(k-3)$ and $\text{E}(\sigma^2 \mid \underset{\sim}{y_s}, k) = (k-1)s_*^2/(k-3),\ k \ge 4$.

For Equation (A6), using the independence of μ and α,

$$E(\bar{Y} \mid \underset{\sim}{y_s}) = E\{\lambda\bar{y} + (1-\lambda)\mu \mid \underset{\sim}{y_s}\} = E(\lambda \mid k)\bar{y} + \{1 - E(\lambda \mid k)\}\bar{y}_*, \tag{A8}$$

where $\lambda = n(\alpha+N)/N(\alpha+n)$ as in Theorem 2. Next, we find V_1 and V_2 in Equation (A7).

First, we find V_1 in (A7). It is easy to show that

$$V_1 = (n-1)(1-f)\frac{s^2}{n}E(\lambda\phi \mid \underset{\sim}{y}_*, k) + E\left\{\lambda(1-\lambda)\phi(\mu-\bar{y})^2 + \lambda(1-\lambda)(1-\phi)\ \frac{\sigma^2}{n}\middle| \underset{\sim}{y}_s, k\right\},$$

where $\phi = 1/(\alpha+n+1)$ as in Theorem 2. Now because α and μ are independent,

$$E\{\lambda(1-\lambda)\phi(\mu-\bar{y})^2 \mid \underset{\sim}{y}_s, k\} = \{(\bar{y}-\bar{y}_*)^2 + \mathrm{Var}(\mu \mid \underset{\sim}{y}_s, k)\}E\{\lambda(1-\lambda)\phi \mid \underset{\sim}{y}_s, k\}.$$

Because $E(\sigma^2 \mid \underset{\sim}{y}_*) = (k-1)s_*^2/(k-3)$ and $\mathrm{Var}(\mu \mid \underset{\sim}{y}_s, k) = (k-2)s_*^2/k(k-3),\ k \geq 4$, we have

$$\begin{aligned} V_1 = {} & (n-1)(1-f)\frac{s^2}{n}E(\lambda\phi \mid k) \\ & + \left\{(\bar{y}-\bar{y}_*)^2 + \frac{(k-2)s_*^2}{k(k-3)}\right\}E\{\lambda(1-\lambda)\phi \mid k\} + \frac{(k-1)s_*^2}{(k-3)n}E\{\lambda(1-\lambda)(1-\phi) \mid k\}. \end{aligned} \tag{A9}$$

Second, we find V_2 in (A7). We use the standard formula for variance,

$$V_2 = E[\{E(\bar{Y} \mid \underset{\sim}{y}_s, \Omega) - E(\bar{Y} \mid \underset{\sim}{y}_s)\}^2]$$

where $E(\bar{Y} \mid \underset{\sim}{y}_s)$ is given by Equation (A8). It is easy to show that

$$E(\bar{Y} \mid \underset{\sim}{y}_s, \Omega) - E(\bar{Y} \mid \underset{\sim}{y}_s) = (\bar{y}-\bar{y}_*)\{\lambda - E(\lambda \mid \underset{\sim}{y}_*, k)\} + (\mu-\bar{y}_*)(1-\lambda).$$

Then, completing the squares and using the independence of μ and α again, we have

$$V_2 = (\bar{y}-\bar{y}_*)^2\mathrm{var}(\lambda \mid k) + \frac{(k-2)s_*^2}{k(k-3)}E\{(1-\lambda)^2 \mid k\}. \tag{A10}$$