Noninformative nonparametric quantile estimation for simple random samples

David Nelson [*]

Center for Chronic Disease Outcomes Research, VA HSR Center of Excellence, Minneapolis, Minnesota

Glen Meeden

University of Minnesota, Minneapolis, Minnesota

Abstract

For noninformative nonparametric estimation of finite population quantiles under simple random sampling, estimation based on the Polya posterior is similar to estimation based on the Bayesian approach developed by Ericson (1969, JRSSB, 31, 195-233) in that the Polya posterior distribution is the limit of Ericson's posterior distributions as the weight placed on the prior distribution diminishes. Furthermore, Polya posterior quantile estimates can be shown to be admissible under certain conditions. We demonstrate the admissibility of the sample median as an estimate of the population median under such a set of conditions. As with Ericson's Bayesian approach, Polya posterior based interval estimates for population quantiles are asymptotically equivalent to the interval estimates obtained from standard frequentist approaches. In addition, for small to moderate populations, Polya posterior based interval estimates for quantiles of a continuous characteristic of interest tend to agree with the standard frequentist interval estimates.

Key words: finite population sampling, admissibility, quantile estimation, noninformative inference, nonparametric inference

1991 MSC: 62D05 Running title: Noninformative Quantile Estimation

1. Introduction

The utility of estimating population quantiles for describing and understanding the distribution of a population characteristic of interest is well recognized.

[*] Corresponding author.
Email addresses: dave.nelson@med.va.gov (David Nelson), glen@stat.umn.edu (Glen Meeden).

Preprint submitted to Elsevier Science *14 June 2004*

We focus here on nonparametric estimation of population quantiles when little or no auxiliary or prior information is available to use in a Bayesian analysis. While prior information is often present in finite population sampling, it is not uncommon to encounter situations in which little or no prior information is available.

The most common approach to estimation in finite population sampling takes a frequentist design based approach, hence a number of design based procedures have been developed for the estimation of population distribution functions and quantiles. Woodruff [14] proposed a method for the construction of point and interval estimates for finite population quantiles based on an inversion of the estimated distribution function. This approach has become the standard in situations lacking auxiliary information or where auxiliary information is built into the sampling design. Francisco and Fuller [5] proposed a general approach to the construction of interval estimates for quantiles under complex survey designs. For simple random samples this approach is similar to the approach proposed by Woodruff.

Ericson [3] developed a noninformative nonparametric proper Bayesian approach to finite population sampling utilizing Dirichlet-multinomial prior distributions. For large populations the credible intervals for estimation of population quantiles generated by this approach are asymptotically equivalent to the intervals obtained from the standard frequentist procedure proposed by Woodruff for simple random samples. Binder [1] and Lo [8] developed more general Dirichlet process extensions of Ericson's approach to finite population sampling. Meeden and Vardeman [12] proposed the application of the noninformative stepwise Bayes Polya posterior for the estimation of finite population quantiles. The Polya posterior is readily applicable in situations where little or no prior auxiliary information is available. Meeden and Vardeman observed that the interval estimates obtained from the Polya posterior performed similarly to the interval estimates obtained from the approach proposed by Woodruff for simple random samples. This similarity between the Polya posterior and frequentist interval estimates was observed in a simulation study examining a number of superpopulation models and in which the form of the Polya posterior interval estimate was approximated using Monte Carlo methods.

Note that when used to estimate the population mean, under squared error loss, the Polya posterior estimate for the mean is the sample mean. The variance of the sample mean under the Polya posterior is asymptotically equivalent to the variance obtained with the standard frequentist asymptotic methods for simple random samples. Polya posterior based estimation of the population mean then is essentially equivalent to estimation of the population mean using the standard frequentist approach.

We investigate similar properties of the conditional predictive distribution for population quantiles generated by the Polya posterior. We focus on the use of the Polya posterior in the estimation of population quantiles in situations consistent with simple random sampling. With a convenient, sensible definition of the population quantile, we demonstrate that Polya posterior based estimation of the population quantile is similar to standard frequentist estimation of the population quantile. The conditional predictive distribution obtained from the Polya posterior is the limit of Ericson's proper posterior distributions, therefore general properties of the Polya posterior and the asymptotic equivalence of Polya posterior interval with those obtained from the frequentist approaches are readily demonstrated by adapting Ericson's results. In addition, under certain conditions, stepwise Bayes properties of Polya posterior quantile estimates can be used to establish the admissibility of these Polya posterior estimates. In particular, we use a simple stepwise Bayes argument to establish the admissibility of the sample median as an estimate for the population median under the absolute error loss function.

2. Samples from finite populations

Consider a finite population consisting of N units labeled 1, 2, ..., N. The labels are assumed to be known but to contain no information about the units. For each unit i let y_i, a real number, be the unknown value of some characteristic of interest. We consider the estimation of quantiles of the state of nature, $\mathbf{y} = (y_1, \ldots, y_N)$. This state of nature is assumed to belong to a set $\mathcal{Y}$, a subset of N-dimensional Euclidean space, $\mathbb{R}^N$. We assume that the values for the characteristic of interest are elements of some specified finite set of k real numbers $\mathbf{b} = \{b_1, \ldots, b_k\}$. The parameter space then will be a subset of

$$\mathcal{Y}(\mathbf{b}) = \{\mathbf{y} \,|\, \text{For } i = 1, \ldots, N,\ y_i = b_j \text{ for some } j = 1, \ldots, k\}.$$

A sample s is a subset of $\{1, 2, \ldots, N\}$ containing $n(s)$ of elements. Let $\mathcal{S}$ denote the set of all possible samples. A sampling design is a probability measure $\mathbf{p}$ defined on $\mathcal{S}$. Given a parameter point $\mathbf{y} \in \mathcal{Y}$ and $s = \{i_1, \ldots, i_{n(s)}\}$, where $1 \leq i_1 < \cdots < i_{n(s)} \leq N$, let $\mathbf{y}_s = \{y_{i_1}, \ldots, y_{i_{n(s)}}\}$. A sample point consists of the set of observed labels s along with the corresponding values for the characteristic of interest, denoted by

$$z = (s,\ y_s = \{y_{s_{i_1}}, \ldots, y_{s_{i_{n(s)}}}\})$$

where we use the notation y_s to distinguish the observed sample values from the values that would be observed for a particular parameter value, $\mathbf{y}_s$. The set of possible sample points depends on both the parameter space and the design. The sample space can be written as

$$Z(\mathcal{Y}(\mathbf{b}),\ \mathbf{p}) = \{(s,\ y_s) \mid \mathbf{p}(s) > 0 \text{ and } y_s = \mathbf{y}_s \text{ for some } \mathbf{y} \in \mathcal{Y}(\mathbf{b})\}.$$

In the following discussion we focus on simple random samples of fixed size n so we suppress the design $\mathbf{p}$ in the notation for the sample space.

3. The Polya posterior

The 'Polya posterior' is a noninformative stepwise Bayes procedure readily used when little or no prior information is available. The Polya posterior is related to the Bayesian bootstrap of Rubin [13]. See Lo [9] for a discussion of the general properties of these methods. Given a data point $z = (s,\ y_s)$ the 'Polya posterior' is a predictive joint distribution for the unobserved units in the population conditional on the observed sample values with an interpretation as a Polya urn sampling scheme for the unobserved elements where the urn initially contains the observed sample. Feller [4] discusses Polya sampling in detail and Meeden and Ghosh [6] discuss the Polya urn interpretation of the Polya posterior in detail.

For a broad range of decision problems, estimation procedures based on the Polya posterior are unique stepwise Bayes estimates and therefore are admissible. Note that a number of standard frequentist estimators are not unique Bayes or generalized Bayes and hence the standard complete class results are not applicable. However, the stepwise Bayes argument has been used to establish the admissibility of a number of estimators [2,7,11,10]. In these stepwise Bayes arguments a finite sequence of disjoint subsets of the finite parameter space is selected and a different prior distribution is defined on each of the subsets. These subsets and priors are considered in order and at each step the Bayes procedure is found for each sample point receiving positive sampling probability under the respective prior distribution which was not considered in earlier steps. This process continues until all possible samples have been considered. The stepwise Bayes estimator is constructed by defining the value of the estimator at a given sample point to be that value identified in the steps described above. If, for all $\mathbf{b}$, the Bayes estimators identified in these steps are unique then the resultant estimator will be admissible. More specifically, for all finite $\mathbf{b}$, the estimator will be admissible for each finite subspace $\mathcal{Y}(\mathbf{b})$ of the parameter space $\mathbb{R}^N$ and therefore the estimator will be finitely admissible. This finite admissibility then yields the admissibility of the estimate. Ghosh and Meeden [6] discuss stepwise Bayes procedures and the relationship between admissibility and finite admissibility.

In the simplest case, applicable to simple random sampling, the stepwise Bayes argument for the Polya posterior partitions the parameter space $\mathcal{Y}(\mathbf{b})$ into subsets $\mathcal{Y}_{\mathbf{b}}(\mathbf{b}')$ where, for $\mathbf{b}' = \{b'_1, \ldots, b'_{k'}\} \subseteq \mathbf{b}$,

$$\mathcal{Y}_{\mathbf{b}}(\mathbf{b}') = \{\mathbf{y} \mid \text{For } i = 1, \ldots, N,\ y_i = b'_j \text{ for some } j = 1, \ldots, k'$$

and for each $j = 1, \dots, k'$ there is an i such that $y_i = b'_j\}$.

For the ordered values b'_1 to $b'_{k'}$, let $c_y(i)$ denote the number of elements in $\mathbf{y}$ taking the value b'_i and let $\Theta_{k'}$ be the $k' - 1$ dimensional simplex. On the subspace $\mathcal{Y}_{\mathbf{b}}(\mathbf{b}')$ we specify the prior distribution

$$\pi\Big(\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}')\Big) \propto \int\limits_{\Theta_{k'}} \prod_{i=1}^{k'} \theta_i^{c_y(i)-1}\, \mathrm{d}\theta.$$

These subspaces and the corresponding prior distributions are considered in lexicographic order of the $\mathbf{b}'$ and are paired with the sample space subspaces

$$Z(\mathcal{Y}_{\mathbf{b}}(\mathbf{b}')) = \{(s,\ y_s) \mid y_s = \mathbf{y}_s \text{ for some } \mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}') \text{ such that for each } j = 1, \dots, k' \text{ there is an } i \text{ such that } y_{s_i} = b'_j\}.$$

For a given parameter point $\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}')$ consistent with $z \in Z(\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'))$ the Polya posterior predictive distribution satisfies

$$\pi\Big(\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}') \mid z \in Z(\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'))\Big) = \frac{\Gamma(n)}{\Gamma(N)} \prod_{i=1}^{k'} \frac{\Gamma(c_y(i))}{\Gamma(c_z(i))}$$

where the number of elements in the sample taking the value b'_i is denoted by $c_z(i)$ for the ordered values b'_1 to $b'_{k'}$. This conditional posterior predictive distribution is equivalent to a Polya urn sampling method as described above.

Since each value in $\mathbf{b}'$ appears at least once in each parameter in $\mathcal{Y}_{\mathbf{b}}(\mathbf{b}')$, the prior distribution specified on this subspace in the stepwise Bayes argument is a proper prior distribution. The Polya posterior predictive distribution for the corresponding pairing of parameter subspace and sample subspace then is a proper posterior distribution. In practice we do not formally specify the entire structure outline above. Given the sample, we simply identify the set of values in the sample, $\mathbf{b}$, and use the Polya posterior predictive distribution,

$$\pi\Big(\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}') \mid z \in Z(\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'))\Big)$$

to make inferences about the population values of the characteristic of interest. While this posterior distribution is not a formal Bayesian posterior distribution we utilize this posterior distribution in a standard Bayesian manner, implicitly working with $\mathcal{Y}(\mathbf{b})$ as the parameter space. For further discussion of the use of these posterior predictive distributions in inference see Ghosh and Meeden [6].

Given a finite set of possible values for the characteristic of interest, many functions of the characteristic, such as the mean and the median, are simply functions of the number of elements taking each of the values. Let $\mathbf{N} = (c_y(1),\ \ldots,\ c_y(k'))$ and let $\mathbf{n} = (c_z(1),\ \ldots,\ c_z(k'))$. The posterior distribution for the counts, $\mathbf{M} = (M_1,\ \ldots,\ M_{k'}) = \mathbf{N} - \mathbf{n} = (c_y(1),\ \ldots,\ c_y(k')) - (c_z(1),\ \ldots,\ c_z(k'))$, for the number of unobserved elements taking each of the values in $\mathbf{b}'$, satisfies

$$\pi\Big(\mathbf{M} \mid z \in Z(\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'))\Big) = \frac{\Gamma(n)}{\Gamma(N)}\ \frac{\Gamma(N-n+1)}{\prod_1^{k'}\Gamma(c_y(i)-c_z(i)+1)}\ \prod_1^{k'}\frac{\Gamma(c_y(i))}{\Gamma(c_z(i))}.$$

Ericson's [3] noninformative nonparametric model is based on the Dirichlet-multinomial prior, $DM(N, \epsilon, \mathbf{e})$,

$$\begin{aligned} P(\mathbf{N}) &= \int_\theta \pi(\mathbf{N} \mid \theta) f(\theta) d\theta \\ &= \int_\theta \Big\{\frac{\Gamma(N+1)}{\prod_1^k \Gamma(c_y(i)+1)} \prod_1^k \theta_{\mathrm{i}}{}^{c_y(i)}\Big\}\ \Big\{\frac{\Gamma(\epsilon)}{\prod_1^k \Gamma(\epsilon_i)} \prod_1^k \theta_i^{\epsilon_i - 1}\Big\}\ \mathrm{d}\theta \end{aligned}$$

for $P(y_j = b_i \mid \theta) = \theta_i$, $\sum_{i=1}^k \theta_i = 1$ with $\mathbf{e} = (\epsilon_1, \ldots, \epsilon_k)$, $\epsilon_i > 0$ for $i = 1, \ldots, k$, and $\epsilon = \sum_1^k \epsilon_i$. Let $c_z(i)$ be the number of sample elements taking value b_i for $i = 1, \ldots, k$, and let $\mathbf{n} + \mathbf{e} = (c_z(1) + \epsilon_1, \ldots, c_z(k) + \epsilon_k)$. The posterior distribution generated by the above prior distribution is the Dirichlet-multinomial $DM(N - n, n + \epsilon, \mathbf{n} + \mathbf{e})$ distribution,

$$P(\mathbf{M} \mid z) = \frac{\Gamma(n+\epsilon)}{\Gamma(N+\epsilon)} \frac{\Gamma(N-n+1)}{\prod_1^k \Gamma(c_y(i)-c_z(i)+1)} \prod_1^k \frac{\Gamma(c_y(i)+\epsilon_i)}{\Gamma(c_z(i)+\epsilon_i)}.$$

Choosing the sum ϵ in the prior distribution for θ to be small incorporates the sense of vagueness or lack of information into the estimation process. Smaller ϵ result in more weight given to the observed data in the form of the posterior distribution. Letting $\epsilon \longrightarrow 0$ the limit of Ericson's posterior distributions, as the weight given to the prior distribution diminishes, is the Polya posterior distribution for the population counts.

4. Quantile estimation with the Polya posterior

Adapting the arguments presented by Ericson we consider Polya posterior based estimation of quantiles of $\mathbf{y}$ for simple random samples of size n. Consistent with Ericson's approach, we define $q_\alpha(\mathbf{y})$, the 100α quantile of $\mathbf{y}$, to be the minimum value such that at least $100\alpha\%$ of the values in $\mathbf{y}$ are less than or equal to this value and at least $100(1-\alpha)\%$ of the values in $\mathbf{y}$ are greater than or equal to this value. This definition restricts the quantile value to the set of values for $\mathbf{y}$ in the finite population and yields a unique value for the population quantile. Furthermore, for a given N and with $\lceil N\alpha \rceil$ the smallest integer not less than $N\alpha$, for $q_\alpha(\mathbf{y})$ to be less than or equal to a value q at least $\lceil N\alpha \rceil$ elements in $\mathbf{y}$ need to be less than or equal to q.

For $i = 1$ to k let $C_z(i) = \sum_1^j c_z(j)$ be the number of sample elements taking values less than or equal to $y_{s(i)}$, the i^{th} value in the ordered set of unique sample values. For $q_\alpha(\mathbf{y})$ to be less than or equal to $y_{s(i)}$ we need at least $\lceil N\alpha \rceil$ of the elements of the population constructed in the Polya urn sampling to be less than or equal to $y_{s(i)}$. Therefore, collapsing the observed sample and the population into two groups based on whether the values are greater than $y_{s(i)}$ or not, the Polya posterior distribution for the quantile can be shown to satisfy

$$\begin{aligned}F_{\alpha,\pi(\cdot\,|\,z)}(y_{s(i)}) &= P_{\pi(\cdot\,|\,z)}(q_\alpha(\mathbf{y}) \le y_{s(i)} \,|\, z) \\ &= \frac{\Gamma(n)}{\Gamma(N)} \sum_{j=\lceil N\alpha \rceil}^{N-n+C_z(i)} \left\{ \frac{\Gamma(j)\,\Gamma(N-j)}{\Gamma(C_z(i))\,\Gamma(n-C_z(i))} \binom{N-n}{j-C_z(i)} \right\}.\end{aligned}$$

Using a similar argument Ericson showed that for the Dirichlet-multinomial posterior distributions

$$P(q_\alpha(\mathbf{y}) \le y_i \mid z) =$$

$$\sum_{j=\lceil N\alpha \rceil}^{N-n+C_z(i)} \binom{N-n}{j-C_z(i)} \frac{\Gamma(n+\epsilon)\Gamma(j+\epsilon_i))\Gamma(N+\epsilon-j+\epsilon_i))}{\Gamma(N+\epsilon)\Gamma(C_z(i)+\epsilon_i)\Gamma(n+\epsilon-C_z(i)+\epsilon_i))}.$$

Again, letting $\epsilon \longrightarrow 0$ the Polya posterior distribution is the limit of Ericson's posterior distribution.

The relatively simple expression for the posterior predictive distribution function for the population quantile can be used to find estimates for the population median for given loss functions. In particular, for certain loss functions and for certain conditions we can show that the estimates derived from this distribution are stepwise Bayes and, hence, are admissible estimates of the population quantile. For example, under squared error loss the mean of this posterior predictive distribution for the quantile is the admissible stepwise Bayes estimate of the population quantile. Under absolute error loss if the median of the Polya posterior distribution for the quantile is unique, for all $\mathbf{b}'$, then the median will be the admissible stepwise Bayes estimate of the population median.

5. Admissibility of the sample median

The Polya posterior can be used to establish the admissibility of the sample median as an estimate of the population median under simple random sampling and certain conditions on the population size and the sample size. Define the sample median in a manner consistent with the population median. In the finite problems considered here a unique median is the unique Bayes estimate for the population median under the absolute error loss function.

Theorem 1 *Consider a simple random sample of size n from a finite population of size N where for each population unit, i, y_i is the unknown value of some real valued characteristic of interest. If either n is odd or N is even then the sample median is an admissible estimate of the population median for the absolute error loss function.*

This result follows from the following Theorem establishing the uniqueness of the median for these cases and, hence, the finite admissibility of the sample median. Finite admissibility with respect to the parameter space $\mathcal{Y}(\mathbf{b})$ for all $\mathbf{b}$ then yields the admissibility of the sample median for the parameter space $\mathbb{R}^N$. Details on the proofs of these theorems are provided in the Appendix.

Theorem 2 *Consider a simple random sample of size n from a finite population of size N where for each population unit, i, y_i is the unknown value of some real valued characteristic of interest. Assume that the unknown state of nature $\mathbf{y}$ is an element of $\mathcal{Y}(\mathbf{b})$ for $\mathbf{b} = \{b_1, \ldots, b_k\}$. If n is odd or if N is even then the sample median is the unique median of the Polya posterior predictive distribution for the median of $\mathbf{y}$.*

Similar sets of conditions on the sample size and the population size can be identified to establish the admissibility of other sample quantiles as estimators of the corresponding population quantiles.

6. Polya posterior quantile interval estimation

We can construct simple (1-α^*) credible intervals for the population quantile using lower and upper bounds defined by the Polya posterior;

$$\max\Big\{y_{s(i)} \,|\, F_{\alpha,\pi(\cdot\,|\,z)}(y_{s(i)}) \leq \frac{\alpha^*}{2}\Big\},\ \ \min\Big\{y_{s(i)} \,|\, F_{\alpha,\pi(\cdot\,|\,z)}(y_{s(i)}) \geq 1 - \frac{\alpha^*}{2}\Big\}.$$

This interval will have posterior probability somewhat greater than 1-α^* under the Polya posterior. In a simulation study examining estimation of the population median, Meeden and Vardeman [12] used Monte Carlo sampling techniques to approximate these Polya posterior interval and observed that these Polya posterior credible sets performed similarly to the standard frequentist confidence intervals proposed by Woodruff in terms of interval length and coverage. The asymptotic form of the distribution for the population quantile generated by the Polya posterior identified below more formally establishes the large sample equivalence of these approaches for general population quantiles.

Theorem 3 *Consider a simple random sample of size n from a finite population of size N in $\mathcal{Y}_{\mathbf{b}}(\mathbf{b}')$. For a sample $z \in Z(\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'))$, as $N \longrightarrow \infty$,*

$$P_{\pi(\cdot\,|\,z)}\Big(q_\alpha(\mathbf{y}) = y_{s(i)}\,|\,z\Big) \;\longrightarrow\; \sum_{k=C_z(i-1)}^{C_z(i)-1} \binom{n-1}{k} (\alpha)^k (1-\alpha)^{n-1-k}.$$

A proof is presented in the Appendix. The same asymptotic form is found for interval estimates based on the proper Bayesian approach of Ericson.

Given this result, for a sufficiently large sample size and a sufficiently large population size, the Polya posterior based distribution for the population quantile will be well approximated by a normal distribution,

$$P_{\pi(\cdot\,|\,z)}\Big(q_\alpha(\mathbf{y}) \leq y_{s(i)}\,|\,z\Big) \approx \Phi\Big(C_z(i) - 1\,|\,(n-1)\alpha,\ (n-1)\alpha(1-\alpha)\Big)$$

where $\Phi(\cdot\,|\,\mu,\ \sigma^2)$ is the normal cumulative distribution function with mean μ and variance σ^2. Hence, the $100(1-\alpha^*)$ credible interval will be well approximated by the interval,

$$\hat{F}^{-1}\Big(\alpha - z_{1-\frac{\alpha^*}{2}}\sqrt{\frac{\alpha(1-\alpha)}{n-1}}\ \Big),\ \ \tilde{F}^{-1}\Big(\alpha + z_{1-\frac{\alpha^*}{2}}\sqrt{\frac{\alpha(1-\alpha)}{n-1}}\ \Big)$$

where $\hat{F}$ is the empirical likelihood function, and

$$\hat{F}^{-1}(\theta) = inf\{x\,|\,\hat{F}(x) \geq \theta\}, \quad \tilde{F}^{-1}(\theta) = sup\{x\,|\,\hat{F}(x) \leq \theta\}.$$

This interval is similar in structure to the approximate confidence interval proposed by Woodruff [14]. Confidence intervals based on Woodruff's approach take the form

$$\hat{F}^{-1}\Big(\alpha - z_{1-\frac{\alpha^*}{2}}\tilde{s}\ \Big),\ \ \hat{F}^{-1}\Big(\alpha + z_{1-\frac{\alpha^*}{2}}\tilde{s}\ \Big)$$

where $\tilde{s}^2 = \frac{N-n}{N-1}n^{-1}\alpha(1-\alpha)$. This approach slightly differs from the credible set approach in the method for identifying the upper bound but for n sufficiently large, with $\frac{n}{N}$ small, these two intervals will tend to be identical. Moreover, the approach can be modified to identify the upper bound in a similar manner to that specified for the Polya posterior interval estimate. Asymptotically, the point and interval estimates generated by the Polya posterior are equivalent to the estimates obtained from the standard frequentist approach to quantile estimation. In addition, for ϵ sufficiently small, these estimates will be the same as the estimates obtained from Ericson's formal Bayesian approach.

For continuous characteristics of interest in small to moderate sized populations the Polya posterior distribution still yields estimates similar to the standard frequentist estimates. The degree of similarity depends upon the quantile considered, the population size, and the sample size. For a population of 1000 elements and a sample containing distinct values, Figure 1 presents the median of the Polya posterior, the sample quantile, and both the Polya posterior and frequentist interval estimates for specific combinations of quantiles and sample sizes. The upper bounds for the both interval estimates were constructed using $sup\{x \,|\, \hat{F}(x) \leq \theta\}$. For quantiles close to the median the point and interval estimates tend to be the same and the Polya posterior interval estimates occasionally span a smaller set of sample elements than the frequentist estimates. For quantiles further from the median small shifts in the ordered sample elements used to form the point and interval estimates occasionally occur with the Polya posterior point estimates and the span of the Polya posterior interval estimates shifted slightly towards the center of the sample.

Place Figure 1 about here

7. Simulation study

We conducted a small simulation study to further examine the similarity and the performance of Polya posterior based quantile estimation relative to standard frequentist quantile estimation. We present the findings for estimating the median and lower decile for small to moderate populations. We considered four simple superpopulation distributions representing variations in the degree of continuity and the degree of skewness of the characteristic of interest within the population. These distributions comprised the Normal distribution, $N(35,\, \sigma{=}5)$, the binomial distribution $Bin(40,\, 0.5)$, the Gamma distribution, $G(4,\, 1)$, and this same Gamma distribution to one significant digit, denoted $DG1(4,\, 1)$. The distribution $DG1(4,\, 1)$ is less discrete than the $Bin(40,\, 0.5)$ distribution. We considered population sizes of 500, 1000, and 2500 with corresponding sample sizes of 25, 40, and 50. For each simulation we constructed 500 populations from the specified superpopulation distribution and for a simple random sample from the population constructed the frequentist and Polya posterior based estimates for the quantile of interest. We considered both the mean and the median of the Polya posterior distribution for the quantile as point estimates of the population quantile.

Although absolute error is the usual loss function when estimating the population median one could use squared error loss. Under this loss function the mean of the Polya posterior distribution is a stepwise Bayes estimator and must be admissible by the same argument used to prove the admissibility of the sample median under absolute error loss. Given a sample one can use the expression for our stepwise Bayes posterior distribution to calculate the value of this estimator. Meeden and Vardeman [12] and Ghosh and Meeden [6], section 2.8.2, present the results of simulation studies comparing the performance of these estimators in estimating the median. They found that for many distributions likely to be encountered in practice the mean of the posterior distribution performed somewhat better than the sample median under both loss functions. This suggests that the sample median need not be the routinely used estimator in practice and that the mean of the Polya posterior for the population median is an attractive alternative to the sample median.

Table 1 presents the average absolute error and root mean square error of the point estimates and the average coverage and length of the interval estimates for the simulations examining estimation of the population median. For the relatively continuous distributions the mean of the Polya posterior distribution yielded better performing point estimates than the sample median and, as indicated by Figure 1, the Polya posterior interval estimates were the same or shorter than the frequentist interval estimates although the coverage did not seem to suffer greatly. For the highly discrete binomial distribution, the standard frequentist approach yielded shorter interval estimates and the mean of the Polya posterior did not exhibit the robustness to the loss function exhibited for the continuous distributions. We found similar results when we considered estimation of the population quartiles.

Place Table 1 about here

Table 2 presents the results found for estimation of the first decile of the characteristic of interest. Here, for the more continuous distributions, the Polya posterior based interval estimates were the same or longer than the frequentist estimates. For certain combinations of population size and sample size the sample decile is not the median of the Polya posterior distribution for the decile. This shift in the sample value used as the point estimate yielded mixed results for the relative performance of the Polya posterior and frequentist point estimates. Lastly, in this case the mean of the Polya posterior exhibited the robustness to the loss function noted by Meeden and Vardeman [12] for each of the superpopulation models including the highly discrete binomial distribution.

Place Table 2 about here

8. Final remarks

In estimating the mean of the population under squared error loss the sample mean can be shown to be the expected value of the Polya posterior distribution for the population mean. However, explicit forms for the Polya posterior distribution of the mean and interval estimates can not easily be constructed and must be estimated through simulation. Interestingly, using this convenient yet sensible definition of the population quantile, we easily can find explicit forms for the Polya posterior predictive distribution for population quantiles as well as for the Polya posterior point and interval estimates.

The stepwise Bayes argument generating the Polya posterior follows a framework similar to the framework Ericson developed for noninformative nonparametric Bayesian inference in finite populations. Consequently, while not a formal Bayesian posterior distribution, the Polya posterior predictive distribution can be thought of as the limit of Ericson's noninformative nonparametric posterior distributions when the weight given to the prior distribution becomes increasingly small, resulting in greater weight being given to the observed data in the structure of the posterior distributions. This stepwise Bayes approach then provides a ready framework for implementing noninformative nonparametric Bayesian inference for finite population sampling that mitigates the prior specification of the set of possible values for the characteristic of interest and the subjective nominal parameter, ϵ, for the Dirichlet prior distribution necessary in Ericson's proposed noninformative framework.

Lastly, Polya posterior estimators tend to possess attractive frequentist properties. For simple random samples, quantile estimates from the Polya posterior can often be shown to be admissible and asymptotically equivalent to the standard frequentist estimates. For small to moderate population sizes and for characteristics of interest that are not highly discrete the Polya posterior estimates tend to mirror the standard frequentist estimates and, if we consider the use of the mean of the Polya posterior distribution as a point estimator, can offer better performance than the standard frequentist estimator.

Appendix

Proof of Theorem 1. This theorem follows from Theorem 2 and the following standard result. Let m_y denote the population median and let m_z be the unique median under $\pi(\cdot \mid z)$. Let a be another estimate for the population median. For convenience, assume $a > m_z$. Then, since $P_{\pi(\cdot \mid z)}(m_y \leq m_z) > \frac{1}{2}$,

$$\begin{aligned} E_{\pi(\cdot \mid z)}\{|m_z - m_y| - |a - m_y|\} &\leq (m_z - a)\ P_{\pi(\cdot \mid z)}(m_y \leq m_z) \\ &\quad + (a - m_z)\ P_{\pi(\cdot \mid z)}(m_y > m_z) \\ &< \frac{1}{2}(m_z - a)\ +\ \frac{1}{2}(a - m_z). \end{aligned}$$

Proof of Theorem 2. Without loss of generality consider the case where the sample values are distinct. Consider the case where n is odd and N is even. The sample median is $y_{s(\frac{n+1}{2})}$ and $\lceil N\alpha \rceil$ is $\frac{N}{2}$. Then

$$P\left(q_{.5}(\mathbf{y}) \leq y_{s(\frac{n+1}{2})} \,|\, z\right) = \frac{\Gamma(n)}{\Gamma(N)} \sum_{\frac{N}{2}}^{N-\frac{n-1}{2}} \left\{ \frac{\Gamma(j)\,\Gamma(N-j)}{\Gamma(\frac{n+1}{2})\,\Gamma(\frac{n-1}{2})} \binom{N-n}{j-\frac{n+1}{2}} \right\},$$

$$P\left(q_{.5}(\mathbf{y}) \geq y_{s(\frac{n+1}{2})} \,|\, z\right) = \frac{\Gamma(n)}{\Gamma(N)} \sum_{\frac{N}{2}+1}^{N-\frac{n-1}{2}} \left\{ \frac{\Gamma(j)\,\Gamma(N-j)}{\Gamma(\frac{n+1}{2})\,\Gamma(\frac{n-1}{2})} \binom{N-n}{j-\frac{n+1}{2}} \right\}$$

Thus $P(q_{.5}(\mathbf{y}) \leq y_{s(\frac{n+1}{2})} \,|\, z) > 0.5$ since the sum of these two probabilities must be greater than one and the first is greater than the second. Furthermore, with

$$P\left(q_{.5}(\mathbf{y}) < y_{s(\frac{n+1}{2})} \,|\, z\right) = P\left(q_{.5}(\mathbf{y}) \leq y_{s(\frac{n-1}{2})} \,|\, z\right)$$

and shifting the index in the second probability above we have that

$$P\left(q_{.5}(\mathbf{y}) < y_{s(\frac{n+1}{2})} \,|\, z\right) = \frac{\Gamma(n)}{\Gamma(N)} \sum_{\frac{N}{2}}^{N-\frac{n+1}{2}} \left\{ \frac{\Gamma(j)\,\Gamma(N-j)}{\Gamma(\frac{n-1}{2})\,\Gamma(\frac{n+1}{2})} \binom{N-n}{j-\frac{n-1}{2}} \right\}$$

and

$$P\left(q_{.5}(\mathbf{y}) \geq y_{s(\frac{n+1}{2})} \,|\, z\right) = \frac{\Gamma(n)}{\Gamma(N)} \sum_{\frac{N}{2}}^{N-\frac{n+1}{2}} \left\{ \frac{\Gamma(j+1)\,\Gamma(N-j-1)}{\Gamma(\frac{n+1}{2})\,\Gamma(\frac{n-1}{2})} \binom{N-n}{j-\frac{n-1}{2}} \right\}.$$

For $2j \geq N,\ \Gamma(j+1)\,\Gamma(N-j-1) - \Gamma(j)\,\Gamma(N-j)$ is nonnegative and so the second probability is greater than the first probability. Hence $P(q_{.5}(\mathbf{y}) \geq y_{s(\frac{n+1}{2})} \,|\, z) > 0.5$. Therefore $y_{s(\frac{n+1}{2})}$ is the unique median of the Polya posterior predictive distribution for the population median. The proofs for the remaining two cases are similar.

Given the potential for multiple medians for the posterior predictive distribution when n is even and N is odd, arguments similar to those above can not be used to establish the admissibility of the sample median in this remaining case.

Proof of Theorem 3. This result is based on the well known result that, for fixed n and $C_z(i)$, the proportion taking a given value in the Polya urn scheme with two values approaches a beta distribution, and more generally that the distribution for the proportions of each type in a Polya urn sampling scheme approaches a Dirichlet distribution. Specifically,

$$P_{\pi(\cdot\,|\,z)}\left(q_\alpha(\mathbf{y}) \leq y_{s(i)} \,|\, z\right) \longrightarrow \frac{\Gamma(n)}{\Gamma(C_z(i))\,\Gamma(n-C_z(i))} \int_\alpha^1 \theta^{C_z(i)-1}(1-\theta)^{n-C_z(i)-1}\,\mathrm{d}\theta$$

as $N \longrightarrow \infty$. If the sample values are distinct integration by parts yields

$$P_{\pi(\cdot\,|\,z)}(q_\alpha(\mathbf{y}) = y_{s(i)}\,|\,z) \longrightarrow \binom{n-1}{C_z(i)-1}(\alpha)^{C_z(i)-1}(1-\alpha)^{n-C_z(i)}.$$

Repeated application of integration by parts yields the general result.

References

[1] D. A. Binder, Non-parametric Bayesian models for samples from finite populations, Journal of the Royal Statistical Society, Series B 44 (1982) 388–393.

[2] L. D. Brown, A complete class theorem for statistical problems with finite sample spaces, Annals of Statistics 9 (1981) 1289–1300.

[3] W. A. Ericson, Subjective Bayesian models in sampling finite populations, Journal of the Royal Statistical Society, Series B 31 (1969) 195–233.

[4] W. Feller, An Introduction to Probability Theory and Its Applications, Volume I, Wiley, New York, 1968.

[5] C. A. Francisco, W. A. Fuller, Quantile estimation with a complex survey design, Annals of Statistics 19 (1991) 454–469.

[6] M. Ghosh, G. Meeden, Bayesian Methods for Finite Population Sampling, Chapman and Hall, London, 1997.

[7] F. C. Hsuan, A stepwise Bayesian procedure, Annals of Statistics 7 (1979) 860–868.

[8] A. Lo, Bayesian statistical inference for sampling a finite population, Annals of Statistics 14 (1986) 1226–1233.

[9] A. Lo, A Bayesian bootstrap for a finite population, Annals of Statistics 16 (1988) 1684–1695.

[10] G. Meeden, M. Ghosh, S. Vardeman, Some admissible nonparametric and related finite population sampling estimators, Annals of Statistics 13 (1985) 811–817.

[11] G. Meeden, M. Ghosh, Admissibility in finite problems, Annals of Statistics 9 (1981) 846–852.

[12] G. Meeden, S. Vardeman, A noninformative Bayesian approach to interval estimation in finite population sampling, Journal of the American Statistical Association 86 (1991) 972–980.

[13] D. Rubin, The Bayesian bootstrap, Annals of Statistics 9 (1981) 130–134.

[14] R. S. Woodruff, Confidence intervals for medians and other position measures, Journal of the American Statistical Association 47 (1952) 635–646.

Table 1
Median Estimation Simulation Results

Population	Polya Posterior (PP) Estimates						Frequentist Estimates			
	PP Mean		PP Median							
	AAE	RMSE	AAE	RMSE	Coverage	Length	AAE	RMSE	Coverage	Length
$y \sim N(\mu = 35, \sigma = 5)$										
$N = 500, n = 25$	0.911	1.165	0.998	1.271	0.962	5.84	0.998	1.271	0.962	5.84
$N = 1000, n = 40$	0.715	0.909	0.794	1.000	0.928	3.82	0.794	1.000	0.970	4.46
$N = 2500, n = 50$	0.631	0.797	0.706	0.884	0.962	3.56	0.706	0.884	0.962	3.56
$y \sim Bin(40, .5)$										
$N = 500, n = 25$	0.559	0.711	0.574	0.821	1.000	4.39	0.574	0.821	0.998	3.64
$N = 1000, n = 40$	0.449	0.562	0.434	0.680	0.994	3.17	0.434	0.680	0.998	2.77
$N = 2500, n = 50$	0.414	0.526	0.368	0.613	0.998	3.06	0.368	0.613	0.994	2.20
$y \sim Gamma(4, 1)$										
$N = 500, n = 25$	0.325	0.407	0.351	0.436	0.988	2.20	0.351	0.436	0.988	2.20
$N = 1000, n = 40$	0.265	0.332	0.282	0.355	0.936	1.41	0.282	0.355	0.970	1.66
$N = 2500, n = 50$	0.243	0.308	0.258	0.321	0.948	1.34	0.258	0.321	0.948	1.34
$y \sim DG1(4, 1)$										
$N = 500, n = 25$	0.331	0.414	0.348	0.446	0.976	2.21	0.348	0.446	0.976	2.15
$N = 1000, n = 40$	0.269	0.335	0.290	0.369	0.964	1.52	0.290	0.369	0.986	1.73
$N = 2500, n = 50$	0.246	0.311	0.261	0.326	0.970	1.40	0.261	0.326	0.968	1.34

Average Absolute Error (AAE) and Root Mean Square Error (RMSE) of Point Estimates

Average Coverage and Average Length of Interval Estimates

Table 2
First Decile Estimation Simulation Results

Population	Polya Posterior (PP) Estimates						Frequentist Estimates			
	PP Mean		PP Median							
	AAE	RMSE	AAE	RMSE	Coverage	Length	AAE	RMSE	Coverage	Length
$y \sim N(\mu = 35, \sigma = 5)$										
$N = 500, n = 25$	1.167	1.476	1.298	1.678	0.922	5.94	1.298	1.678	0.922	5.94
$N = 1000, n = 40$	0.927	1.151	1.033	1.290	0.968	6.78	1.075	1.357	0.946	6.30
$N = 2500, n = 50$	0.835	1.023	0.922	1.154	0.980	6.84	0.922	1.154	0.980	6.84
$y \sim Bin(40, .5)$										
$N = 500, n = 25$	0.743	0.927	0.816	1.092	0.966	3.72	0.816	1.092	0.966	3.72
$N = 1000, n = 40$	0.583	0.732	0.592	0.841	0.992	4.19	0.638	0.948	0.984	3.89
$N = 2500, n = 50$	0.543	0.671	0.548	0.782	0.998	4.33	0.572	0.820	0.998	4.33
$y \sim Gamma(4, 1)$										
$N = 500, n = 25$	0.264	0.347	0.286	0.372	0.910	1.26	0.286	0.372	0.910	1.26
$N = 1000, n = 40$	0.210	0.270	0.246	0.319	0.972	1.41	0.235	0.296	0.964	1.29
$N = 2500, n = 50$	0.195	0.246	0.217	0.275	0.960	1.34	0.217	0.275	0.960	1.34
$y \sim DG1(4, 1))$										
$N = 500, n = 25$	0.278	0.354	0.308	0.391	0.930	1.27	0.308	0.391	0.930	1.27
$N = 1000, n = 40$	0.212	0.273	0.247	0.324	0.988	1.41	0.238	0.300	0.978	1.29
$N = 2500, n = 50$	0.197	0.249	0.218	0.277	0.978	1.34	0.218	0.279	0.978	1.34

Average Absolute Error (AAE) and Root Mean Square Error (RMSE) of Point Estimates

Average Coverage and Average Length of Interval Estimates

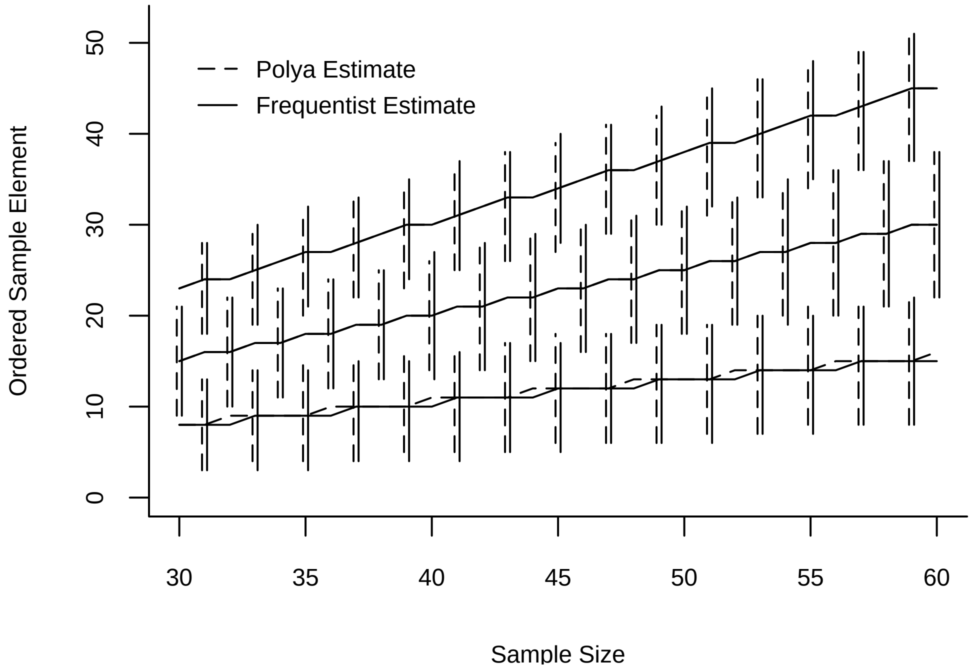

Fig. 1. Polya Posterior and Frequentist Point and Interval Estimates for Lower Quartile, Median, and Upper Quartile: Curves and vertical segments identify which elements of the sample order statistic are identified as the point estimate and the bounds of the interval estimate, respectively.