

Polya Posterior Quantile Estimation for Stratified Populations

David Nelson

VA HSR Center for Chronic Disease Outcomes Research
School of Medicine, University of Minnesota

and

Glen Meeden

School of Statistics, University of Minnesota

Abstract: We develop extensions of the Polya posterior that can be used to estimate population quantiles in stratified populations. In the simple case with no auxiliary information, estimates based on this extension to the Polya posterior perform similarly to the standard frequentist estimates. The Polya posterior can be extended to incorporate a variety of prior knowledge about an auxiliary characteristic. In particular, we consider estimation of population quantiles for stratified populations when there exists prior knowledge that a population quantile of an auxiliary variable lies in a specified interval. This type of prior information is difficult to exploit using standard methods. We show how a constrained version of the Polya posterior can use this information to obtain improved point and interval estimators of a population quantile.

1. Introduction

In the area of finite population sampling more effort has been spent on the development of methods for estimating population means than on developing methods for estimating population quantiles. This imbalance is particularly great for populations with an auxiliary characteristic. In addition, for populations with an auxiliary variable it is usually assumed that the population mean of the auxiliary characteristic is known precisely, a population quantile of the auxiliary characteristic is known precisely, or that there exists some linear relationship between the characteristic of interest and the completely known auxiliary characteristic. We consider estimation of population quantiles for stratified populations when there exists a priori knowledge that a population quantile of an auxiliary variable belongs to a specified interval of real numbers. This type of prior information is difficult to exploit using standard

methods.

The standard Polya posterior is a noninformative, or objective, nonparametric Bayesian posterior predictive distribution which can be used when little or no prior information is available. It is related to the Bayesian bootstrap of Rubin [7], also see Lo [3] for more details. The Polya posterior has been extended to a variety of situations involving different types of prior information. For a broader discussion of the Polya posterior than is presented here see Ghosh and Meeden [2]. Nelson and Meeden [5] examined the use of a constrained version of the Polya posterior to estimate population means and medians for nonstratified populations when, a priori, it is known that the median of an auxiliary characteristic belongs to some specified interval. Nelson and Meeden [6] also examined the form of the Polya posterior predictive distribution for population quantiles for nonstratified populations.

We extend these previous results to consider estimation of a population quantile for a stratified population when it is known a priori that a population quantile of an auxiliary characteristic belongs to a specified interval and related types of prior information. We show how a constrained version of the Polya posterior can use this information to obtain sensible point and interval estimators of a population quantile. The approach developed is applicable to an array of situations broader than knowledge that the particular quantile belongs to a specified interval.

In the following section we introduce the necessary notation for discussing sampling from stratified populations. We then briefly review the Polya posterior and discuss its application to quantile estimation in stratified populations. Subsequently we develop a constrained Polya posterior for estimating population quantiles in stratified populations when prior information about an auxiliary characteristic is available. We present the results of simulation studies examining the performance of median estimators based on this constrained Polya posterior using prior information that the population median of an auxiliary variable is known to belong to a specified interval.

2. Stratified Finite Population Samples

Consider a finite population consisting of G strata with N_j units within each stratum, labeled j_1, j_2, \dots, j_{N_j} for $j = 1 \dots, G$. We assume that the labels are known but contain no information about the values of the characteristics for the units. For each unit j_i , y_{j_i} is the unknown real-valued measure of some characteristic of interest and x_{j_i} is the unknown real-valued measure of an auxiliary characteristic, if present. We assume the stratum membership of each element of the population is known. We consider estimation of quantiles of the characteristic of interest where the state of nature

$$\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_G), \text{ with } \mathbf{y}_j = \left((y_{j_1}, x_{j_1}), \dots, (y_{j_{N_j}}, x_{j_{N_j}}) \right),$$

belongs to a subset of $\mathbb{R}^N \otimes \mathbb{R}^N$ for $N = \sum_1^G N_j$. We assume that the elements of \mathbf{y}_j belong to some specified finite collection of G sets of real numbers $\mathbf{b}_j = \{b_{j_1}, \dots, b_{j_{k_j}}\}$

so the parameter space, $\mathcal{Y}(\mathbf{b}_1, \dots, \mathbf{b}_G)$, can be expressed as

$$\left\{ \mathbf{y} \mid \text{For } j = 1, \dots, G \text{ and } i = 1, \dots, N_j, (y_{j_i}, x_{j_i}) = b_{j_k} \in \mathbf{b}_j \text{ for some } k \right\}.$$

A sample s is a subset of

$$\bigcup_{j=1}^G \{j_1, \dots, j_{N_j}\}$$

containing $n(s) = \sum_1^G n_j(s)$ elements where $n_j(s)$ is the number of sampled elements from the j^{th} stratum. Stratified sampling defines a probability measure \mathbf{p} on \mathcal{S} , the set of all possible samples. For a parameter point $\mathbf{y} \in \mathcal{Y}(\mathbf{b}_1, \dots, \mathbf{b}_G)$ and a sample $s = \bigcup_1^G \{j_{i_1}, \dots, j_{i_{n_j(s)}}\}$, where $j_1 \leq j_{i_1} < \dots < j_{i_{n_j(s)}} \leq j_{N_j}$, define

$$\mathbf{y}_s = \bigcup_1^G \left((y_{j_{i_1}}, x_{j_{i_1}}), \dots, (y_{j_{i_{n_j(s)}}}, x_{j_{i_{n_j(s)}}}) \right).$$

A sample point,

$$z = (s, y_s) = \bigcup_1^G \left((j_{i_1}, y_{j_{i_1}}, x_{j_{i_1}}), \dots, (j_{i_{n_j(s)}}, y_{j_{i_{n_j(s)}}}, x_{j_{i_{n_j(s)}}}) \right),$$

consists of the set of observed labels s along with the corresponding values for the characteristic of interest and the auxiliary characteristic, if present. The set of possible sample points then depends on both the parameter space and the design. The sample space is then given by

$$Z\{\mathcal{Y}(\mathbf{b}_1, \dots, \mathbf{b}_G), \mathbf{p}\} = \left\{ (s, y_s) \mid \mathbf{p}(s) > 0 \text{ and } y_s = \mathbf{y}_s \text{ for some } \mathbf{y} \in \mathcal{Y}(\mathbf{b}_1, \dots, \mathbf{b}_G) \right\}.$$

In what follows we focus on stratified samples comprising simple random samples of specified fixed size, n_j , within each stratum so, for convenience, we suppress the design \mathbf{p} and use the notation $Z(\mathcal{Y}(\mathbf{b}_1, \dots, \mathbf{b}_G))$ for the sample space.

3. A Polya Posterior for Stratified Sampling

The standard ‘Polya posterior’ is derived from a noninformative nonparametric stepwise Bayes estimation procedure. Stepwise Bayes arguments proceed by specifying a finite sequence of disjoint subsets of the finite parameter space with a different prior distribution defined on each of the subsets. These subsets and priors are considered in order where 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 for a given sample point is defined to be the value of the estimator identified in the step in which that sample point

was considered in the above process. If, for all \mathbf{b} , the Bayes estimators identified in these steps are unique then the resultant estimator will be admissible.

Under a particular stepwise Bayes argument, given an observed sample $z = (s, y_s)$, the ‘Polya posterior’ is the predictive posterior distribution for the unobserved units in the population conditional on the observed sample values derived from that sample in the stepwise Bayes argument. This posterior distribution is equivalent to a Polya urn sampling distribution for the unobserved elements where the urn initially contains the observed sample. Feller [1] discusses Polya sampling in detail, Ghosh and Meeden [2] present a detailed discussion of stepwise Bayes theory, the Polya posterior, and the Polya urn interpretation of the Polya posterior.

Here we develop an extension of the Polya posterior applicable to samples from stratified populations. In this and the following sections we omit the auxiliary characteristic for convenience. The stepwise Bayes argument for this modification of the Polya posterior partitions the parameter space $\mathcal{Y}(\mathbf{b}_1, \dots, \mathbf{b}_G)$ into subsets

$$\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G) \text{ where } \mathbf{b}'_j = \{b'_{j_1}, \dots, b'_{j_{k'_j}}\} \subseteq \mathbf{b}_j.$$

For a given $\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G)$, within each stratum the y_{j_i} only take values in \mathbf{b}'_j and for each value in \mathbf{b}'_j there is at least one element in the stratum for which y_{j_i} takes that value. On the subspace $\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G)$ we specify the prior distribution

$$\pi(\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G)) \propto \prod_{j=1}^G \left\{ \int_{\Theta_{k'_j}} \prod_{j_i=1}^{j_{k'_j}} \theta_{j_i}^{c_{\mathbf{y}}(i:j)-1} d\theta \right\}$$

where $\Theta_{k'_j}$ is the $k'_j - 1$ dimensional simplex and, for the ordered values in \mathbf{b}'_j , $c_{\mathbf{y}}(i : j)$ denotes the number of elements in the j^{th} stratum taking the value b'_{j_i} .

These parameter subspaces are paired with the similarly defined sample subspaces $Z(\mathcal{Y}(\mathbf{b}'_1, \dots, \mathbf{b}'_G))$ and are considered in lexicographic order of the $\mathbf{b}'_1, \dots, \mathbf{b}'_G$. For a given parameter point $\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G)$ consistent with an observed sample $z \in Z(\mathcal{Y}(\mathbf{b}'_1, \dots, \mathbf{b}'_G))$ the posterior predictive distribution takes the form

$$\pi(\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G) \mid z \in Z(\mathcal{Y}(\mathbf{b}'_1, \dots, \mathbf{b}'_G))) \propto \prod_{j=1}^G \left\{ \prod_{i=1}^{k'_j} \frac{\Gamma(c_{\mathbf{y}}(i : j))}{\Gamma(c_z(i : j))} \right\}$$

where the number of elements in the sample falling in the j^{th} stratum and taking the value b'_{j_i} is denoted by $c_z(i : j)$. This posterior predictive distribution is equivalent to independent Polya urn sampling within each of the strata.

In practice, given the sample, we simply identify the sets of values observed within each stratum, $\{\mathbf{b}'_1, \dots, \mathbf{b}'_G\}$, along with the respective counts within each stratum, namely the $(c_z(i : j))$, and use the informal yet proper Polya posterior predictive distribution,

$$\pi(\mathbf{y} \in \mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G) \mid z \in Z(\mathcal{Y}(\mathbf{b}'_1, \dots, \mathbf{b}'_G)))$$

to make inferences about the population. While this Polya posterior distribution is not a formal Bayesian posterior distribution we utilize this posterior distribution in a standard Bayesian manner, implicitly working with $\mathcal{Y}_{\mathbf{b}}(\mathbf{b}'_1, \dots, \mathbf{b}'_G)$ as the parameter space.

For quantities such as the mean of the characteristic of interest and, as discussed below, the quantiles of the characteristic of interest a closed form can be found for the point estimator generated by this Polya posterior distribution under certain loss functions. For other quantities it can be difficult to identify a closed form for the estimator but in these cases the above posterior distribution can be used in a Monte Carlo estimation procedure to approximate the Polya posterior estimator. Specifically, given the sample, pseudo-populations can be generated by Polya urn sampling values for the unobserved population elements using the observed sample elements as the initial values in the urns. For each pseudo-population we can construct the relevant quantity of interest and use, say under squared error loss, the mean of these values from many replications of this process as an approximation to the Polya posterior estimator. In addition, we can form interval estimates for the quantity of interest from these repeated draws from the Polya posterior. A simple $1 - 2\alpha$ credible set for the characteristic of interest can be formed from the α and $1 - \alpha$ quantiles of these values for the characteristic of interest in the pseudo-populations.

4. Quantile Estimation in Stratified Populations

Recently, Nelson and Meeden [6] examined the form of the predictive distribution for population quantiles generated by the standard Polya posterior. The distribution considered there is applicable to simple random samples and situations where little prior information is available. We modify the argument presented there to identify the distribution of the population quantiles in a stratified population based on the extension of the Polya posterior given above.

Given a finite set of possible values for the characteristic of interest, many functions of the characteristic, such as the mean and the quantiles of the characteristic of interest, are simply functions of the number of elements taking each of the values. The Polya posterior distributions generates a posterior distribution for these counts. Within each stratum denote the counts of the number of unobserved elements taking each of the values in \mathbf{b}'_j by

$$\mathbf{M}_j = (c_y(1 : j) - c_z(1 : j), \dots, c_y(j_{k'_j} : j) - c_z(j_{k'_j} : j)).$$

The Polya posterior based distribution for these counts of the number of unobserved elements satisfies

$$\pi(\mathbf{M}_1, \dots, \mathbf{M}_G | z) = \prod_1^G \left\{ \frac{\Gamma(n_j)}{\Gamma(N_j)} \binom{N_j - n_j}{\mathbf{M}_j} \prod_1^{k'_j} \frac{\Gamma(c_y(i : j))}{\Gamma(c_z(i : j))} \right\}.$$

This posterior distribution can be used to find the Polya posterior based predictive distribution for any quantile of the characteristic of interest.

To develop this posterior distribution we employ the Polya urn interpretation of the Polya posterior distribution categorizing the observed sample values as to whether they fall above or below $y_s(l)$, the l^{th} observed ordered sample value. Consider the cumulative distribution function,

$$P_{\pi(\cdot|z)}(q_\alpha(\mathbf{y}) \leq y_s(l)),$$

where $q_\alpha(\mathbf{y})$ is the α quantile of the population. For the α quantile of \mathbf{y} to be less than or equal to $y_s(l)$ we need at least $\lceil N\alpha \rceil$ of the population values for the characteristic of interest to be less than or equal to $y_s(l)$. Let

$$C_{\mathbf{y}}(l : j) = \sum_{y_{j_i} \leq y_s(l)} c_{\mathbf{y}}(i : j)$$

and

$$C_z(l : j) = \sum_{y_{s_{j_i}} \leq y_s(l)} c_z(i : j)$$

be the number of population elements and sample elements, respectively, in the j^{th} stratum taking a value less than or equal to $y_s(l)$. The posterior cumulative distribution function for the quantile takes the form

$$\begin{aligned} & P_{\pi(\cdot|z)}(q_\alpha(\mathbf{y}) \leq y_s(l)) \\ &= \sum_{\mathbf{C}_{\mathbf{y}}(l, \alpha)} \prod_{j=1}^G \left\{ \frac{\Gamma(C_{\mathbf{y}}(l : j)) \Gamma(N_j - C_{\mathbf{y}}(l : j))}{\Gamma(C_z(l : j)) \Gamma(n_j - C_z(l : j))} \binom{N_j - n_j}{C_{\mathbf{y}}(l : j) - C_z(l : j)} \frac{\Gamma(n_j)}{\Gamma(N_j)} \right\} \end{aligned}$$

where the sum is restricted to the set of population counts

$$\begin{aligned} \mathbf{C}_{\mathbf{y}}(l, \alpha) &= \left\{ (C_{\mathbf{y}}(l : 1), \dots, C_{\mathbf{y}}(l : G)) \mid \right. \\ &\quad \left. C_{\mathbf{y}}(l : j) > C_z(l : j) \forall j, \sum_{j=1}^G C_{\mathbf{y}}(l : j) \geq \lceil N\alpha \rceil \right\} \end{aligned}$$

for which the population quantile is less than $y_s(l)$. With the n_j fixed,

$$\begin{aligned} & P_{\pi(\cdot|z)}(q_\alpha(\mathbf{y}) \leq y_s(l)) \longrightarrow \\ & \int_{\Theta^*} \left\{ \prod_{j=1}^G \frac{\Gamma(n_j)}{\Gamma(C_z(l : j)) \Gamma(n_j - C_z(l : j))} \theta_j^{C_z(l : j) - 1} (1 - \theta_j)^{n_j - C_z(l : j) - 1} \right\} d\theta \end{aligned}$$

as the $N_j \longrightarrow \infty$, where

$$\Theta^* = \left\{ \theta \in (0, 1)^G \mid \alpha \leq \sum_1^G \frac{N_j}{N} \theta_j \right\}.$$

This is the probability that the weighted average of G independent beta random variables, with parameters $(C_z(l : j), n_j - C_z(l : j))$, is at least α for weights given by the proportions of the population within each stratum.

5. The Constrained Polya Posterior

Nelson and Meeden [5] considered an extension of the standard Polya posterior to incorporate types of prior information about the population of interest which restrict the population to subsets of the usual parameter space. Consider two finite parameter spaces \mathcal{Y} and $\mathcal{Y}_A \subset \mathcal{Y}$. If we specify a prior distribution π on \mathcal{Y} we can then specify a related prior distribution π_A on \mathcal{Y}_A defined by

$$\pi_A(y) = \frac{\pi(y)}{\pi(\mathcal{Y}_A)} = \frac{\pi(y)}{\sum_{y \in \mathcal{Y}_A} \pi(y)}$$

for each $y \in \mathcal{Y}_A$. Given this structure, for a $y \in \mathcal{Y}_A$ consistent with an observed sample z

$$\pi_A(y | z) = \frac{\pi_A(y)}{\pi_A(z)} = \frac{\pi(y | z)}{\pi(\mathcal{A} | z)}.$$

If the available prior information about the population restricts the population parameters to an identifiable subset of $\mathcal{Y}(\mathbf{b}_1, \dots, \mathbf{b}_G)$ then, in general, we can develop a stepwise Bayes argument for which the posterior predictive distribution is the Polya posterior restricted to this subset. In practice, if we can not easily identify the resultant form of the posterior distribution for the quantity of interest, we can approximate the distribution by implementing a rejection sampling version of the Monte Carlo estimation process described above in which we only accept those results of the Polya urn sampling process which satisfy the prior information.

A technical detail that needs to be considered in the full stepwise Bayes argument generating this constrained Polya posterior, as discussed in Nelson and Meeden [5], concerns the consistency of the sample with the prior information. Some samples may not be consistent with the prior information in that it may be impossible to construct pseudo-populations from the sample that meet the constraints given by the prior information. For example, if we know the median of an auxiliary characteristic but all of the observed sample values for the auxiliary characteristic fall below this value it is impossible to construct a pseudo-population for which this value will be the median of the auxiliary characteristic. These technical issues for the stepwise Bayes argument can be easily handled as presented in Nelson and Meeden [5]. We focus here on samples which are consistent with the prior information.

6. Knowledge Concerning an Auxiliary Characteristic

Consider estimation of population quantiles for the characteristic of interest when we have prior information concerning the auxiliary characteristic which places a set of constraints on the number of elements of the population taking specific values. More specifically, consider a partition of the possible values for the auxiliary characteristic

given by the collection of sets \mathcal{A}_i , $i = 1, \dots, p$ for some p . Based upon this partition consider a collection of q groupings of these partition sets,

$$\mathcal{B}_k = \left\{ \mathcal{A}_{k_1}, \dots, \mathcal{A}_{k_{B_k}} \right\}, \quad k = 1, \dots, q,$$

together with a collection of q bounds, $\{l_k, u_k\}$, specifying a set of constraints on the auxiliary characteristic given by

$$l_k \leq \sum_{l=1}^{B_k} \sum_{j=1}^G \sum_{x_{j_i} \in \mathcal{A}_{k_l}} c_y(i : j) \leq u_k, \quad k = 1, \dots, q.$$

These constraints then specify that, for each $k = 1, \dots, q$, the number of elements taking values which fall in any of the sets in \mathcal{B}_k lies between l_k and u_k . The results discussed above can be integrated to construct a constrained Polya posterior generated predictive distribution for any quantile of the characteristic of interest that places positive posterior probability only on population parameters that satisfy this set of constraints on the auxiliary characteristic.

With

$$\mathcal{A}_0^l = (-\infty, y_s(l)], \quad \mathcal{A}_1^l = (y_s(l), \infty).$$

define

$$C_{\mathbf{y}}^u(l, v, j) = \sum_{y_{j_i} \in \mathcal{A}_u^l, x_{j_i} \in \mathcal{A}_v} c_y(i : j)$$

to be the number of population elements in the j^{th} stratum with the characteristic of interest in \mathcal{A}_u^l and with the auxiliary characteristic in \mathcal{A}_v . Let $C_{\mathbf{z}}^u(l, v, j)$ be the analogous sample count and let $D^u(l, v, j)$ be the difference between the respective population counts and the sample counts. Let

$$C_{\mathbf{y}, \mathbf{z}}^{N_j - n_j}(l, j) = \left(D^0(l, 1, j), D^0(l, 2, j), \dots, D^1(l, p, j) \right)$$

Define $C_{\alpha, \mathcal{A}}$ to be the set of vectors for the population counts which satisfy the constraints on the auxiliary characteristic, namely

$$l_k \leq \sum_{m=1}^{B_k} \sum_{j=1}^G \sum_{x_{j_i} \in \mathcal{A}_{k_m}} c_y(i : j) = \sum_{m=1}^{B_k} \sum_{j=1}^G \sum_{u=0}^1 C_{\mathbf{y}}^u(l, k_m, j) \leq u_k$$

for all $k = 1, \dots, q$, and let $C_{l, \alpha, \mathcal{A}}$ be the subset of $C_{\alpha, \mathcal{A}}$ containing vectors for the population counts which satisfy

$$\sum_{j=1}^G \sum_{i=1}^p C_{\mathbf{y}}^0(l, i, j) \geq \lceil N\alpha \rceil$$

The set $C_{l, \alpha, \mathcal{A}}$ then identifies the set of population counts satisfying the above constraints which yield a quantile for the characteristic of interest no greater than $y_s(l)$.

Modifying the arguments above yields the posterior predictive distribution

$$\begin{aligned} & P_{\pi(\cdot|z)}(q_\alpha(\mathbf{y}) \leq y_s(l)) \\ &= \frac{\sum_{C_{l,\alpha,\mathcal{A}}} \left\{ \prod_{j=1}^G \left\{ \prod_{u=0}^1 \prod_{v=1}^p \frac{\Gamma(C_{\mathbf{y}}^u(l,v,j))}{\Gamma(C_{\mathbf{z}}^u(l,v,j))} \right\} C_{\mathbf{y},\mathbf{z}}^{N_j-n_j}(l,j) \frac{\Gamma(n_j)}{\Gamma(N_j)} \right\}}{\sum_{C_{\alpha,\mathcal{A}}} \left\{ \prod_{j=1}^G \left\{ \prod_{u=0}^1 \prod_{v=1}^p \frac{\Gamma(C_{\mathbf{y}}^u(l,v,j))}{\Gamma(C_{\mathbf{z}}^u(l,v,j))} \right\} C_{\mathbf{y},\mathbf{z}}^{N_j-n_j}(l,j) \frac{\Gamma(n_j)}{\Gamma(N_j)} \right\}}. \end{aligned}$$

For fixed n_j , as all $N_j \rightarrow \infty$ the cumulative distribution approaches

$$P_{\pi(\cdot|z)}(q_\alpha(\mathbf{y}) \leq y_s(l)) \rightarrow \frac{\int_{\Theta_{\mathcal{A}}^*} \left\{ \prod_{j=1}^G \Gamma(n_j) \prod_{u,v} \frac{\theta_{ju,v}^{C_{\mathbf{z}}^u(l,v,j)-1}}{\Gamma(C_{\mathbf{z}}^u(l,v,j))} \right\} d\theta}{\int_{\Theta_{\mathcal{A}}} \left\{ \prod_{j=1}^G \Gamma(n_j) \prod_{u,v} \frac{\theta_{ju,v}^{C_{\mathbf{z}}^u(l,v,j)-1}}{\Gamma(C_{\mathbf{z}}^u(l,v,j))} \right\} d\theta}$$

where $\Theta_{\mathcal{A}}$ is the set of $\theta = (\theta_{10,1}, \dots, \theta_{G1,p})$, with $\theta_{j_{i'},j'} \in (0, 1)$, which satisfy the conditions

$$\sum_{u,v} \theta_{ju,v} = 1 \quad \text{for all } j = 1, \dots, G,$$

$$\frac{l_k}{N} \leq \sum_{j=1}^G \sum_u \sum_{v \ni \mathcal{A}_v \in \mathcal{B}_k} \theta_{ju,v} \leq \frac{u_k}{N} \quad \text{for each } k = 1, \dots, q$$

and $\Theta_{\mathcal{A}}^*$ is the subset of $\Theta_{\mathcal{A}}$ satisfying the additional condition

$$T = \sum_{j=1}^G \frac{N_j}{N} \left\{ \sum_v \theta_{j0,v} \right\} \geq \alpha.$$

Consider a collection of G independent Dirichlet random variables,

$$\theta_j = (\theta_{j0,1}, \theta_{j0,2}, \dots, \theta_{j1,p}) \sim D(C_{\mathbf{z}}^0(l, 1, j), C_{\mathbf{z}}^0(l, 2, j), \dots, C_{\mathbf{z}}^1(l, p, j)).$$

T is the weighted average of the sums of the first half of the components of the individual Dirichlet random variables, with weights N_j/N . The asymptotic form for the above Polya posterior distribution is equivalent to the probability that T is at least α conditional on the Dirichlet random variables, θ_j , satisfying the constraints directly above.

As an example of the situation we are considering assume we know a priori that the median of the auxiliary characteristic falls in the interval $\mathcal{A} = [a_1, a_2]$. For the auxiliary characteristic median to fall in \mathcal{A} we need at least $\lceil \frac{N}{2} \rceil$ of the population values for the auxiliary characteristic to be less than a_2 and at least $\lceil \frac{N}{2} \rceil$ of the population values for the auxiliary characteristic to be greater than a_1 . Specifically then, with

$$\mathcal{A}_1 = (-\infty, a_1), \quad \mathcal{A}_2 = [a_1, a_2], \quad \mathcal{A}_3 = (a_2, \infty)$$

$$\mathcal{A}_0^l = (-\infty, y_s(l)], \quad \mathcal{A}_1^l = (y_s(l), \infty)$$

we have the constraints

$$\sum_{v=1}^2 \sum_{j=1}^G \sum_{u=0}^1 C_{\mathbf{y}}^i(l, v, j) \geq \left\lceil \frac{N}{2} \right\rceil, \quad \sum_{v=2}^3 \sum_{j=1}^G \sum_{u=0}^1 C_{\mathbf{y}}^i(l, v, j) \geq \left\lceil \frac{N}{2} \right\rceil$$

and for the quantile of the characteristic of interest to be less than $y_s(l)$ we have the condition that

$$\sum_{j=1}^G \sum_{v=1}^3 C_{\mathbf{y}}^0(l, v, j) \geq \lceil N\alpha \rceil.$$

7. Simulation Results

The discussion above focused upon identifying the forms of Polya posterior based posterior predictive distributions for population quantiles applicable to sampling from a stratified population. The Polya urn schemes associated with the identified posterior distributions provide an intuitive basis for developing estimators of the population quantiles. As discussed above, the estimators derived from the Polya posterior techniques can often be shown to be unique stepwise Bayes estimators and hence can be shown to be admissible estimators. Although admissibility is a property that an estimator would be expected to possess, good estimation techniques would be expected to possess additional desirable properties. For instance, it is often of interest to obtain not only point estimates but also interval estimates for the quantity of interest. In addition to point estimates which perform well a good estimation technique also should provide interval estimates which perform well. Polya posterior based estimators have been observed to possess desirable frequentist properties for a variety of estimation problems. We could then expect the point and interval estimates obtained from the Polya posterior distributions developed above to perform well from a frequentist standpoint.

We conducted a small simulation study to investigate the performance of the Polya posterior median estimators for stratified populations. In each simulation, two hundred fifty populations were formed from a specified stratified superpopulation model comprising both a characteristic of interest and an auxiliary characteristic. These superpopulation models are summarized in Table 1. Nelson and Meeden [6] examined the performance of the posterior predictive distribution obtained from the standard Polya posterior. They observed that for continuous characteristics of interest the Polya posterior estimates tended to yield some improvement over the standard frequentist estimators but for highly discrete characteristics of interest the standard frequentist estimators tended to perform better than the Polya posterior. We considered three general classes of superpopulation models for which both characteristics were continuous and three classes for which both characteristics were discrete.

The different model classes are presented in Table 1. Within each class the values for the auxiliary characteristic within each stratum were constructed using different

members of the same family of distributions. The distributions used in the individual strata are listed in the Table. Given the values for the auxiliary characteristic, the characteristic of interest was generated using a conditional linear model distribution where the variance for the conditional distribution varied within the class of simulations. The distributions for the auxiliary characteristic in the last class of superpopulation models presented in Table 1, denoted $U(a : b)$, place a uniform distribution on the integers between a and b .

For each generated population a stratified random sample was drawn, sampling 10% of each stratum population. For each sample we computed the values of the standard frequentist point and 95% confidence interval estimates of the population median using the inversion of the empirical cumulative distribution function as developed by Woodruff [9] and outlined in Särndal, Swensson and Wretman [8]. The empirical cumulative distribution function is the weighted combination of strata specific empirical distribution functions with weights equal to the stratum proportion of the total population.

In addition, we computed the point and interval estimates obtained from two different Polya posterior distributions. The first, as discussed in section 4, did not use any information about the auxiliary characteristic. The second Polya posterior distribution for the population median did use information about the auxiliary characteristic, as discussed in section 6. This Polya posterior distribution considered knowledge that the median of the auxiliary characteristic fell in the interval formed by the 45th and 55th population percentiles of the auxiliary characteristic.

For each approach, Polya posterior based point estimates and the 95% credible set interval estimate for the population median were obtained for each sample. As point estimators we considered both the median of the Polya posterior distribution and the mean of the Polya posterior distribution. Meeden and Vardeman [4], and subsequently Nelson and Meeden [6] observed that the mean of the Polya posterior for a population quantile often performs better than both the median of the posterior distribution and the standard frequentist estimators. We approximated the form of the Polya posterior quantile distributions to construct each of these Polya posterior based estimators. These approximations used Monte Carlo and rejection sampling Monte Carlo methods based on the respective Beta and Dirichlet interpretations for the asymptotic forms of the Polya posterior quantile distributions given above.

Table 1 summarizes the results of the simulations examining the performance of the median of the Polya posterior distributions as point estimates of the population median. A summary of the performance of the interval estimates obtained from the Polya posterior distributions also is presented in this table. For each estimator we computed the mean absolute deviation of the estimates and the average coverage and length of the interval estimates for the 250 replications in the simulation.

For the superpopulation models comprising continuous characteristics the performance of the Polya posterior based estimators which ignores the prior information about the auxiliary characteristic is similar to that of the standard frequentist estima-

Table 1: Summary of Simulation Results: Median Estimation using Median of Polya Posterior

| Population | Polya Estimates using Prior Information | | | Polya Estimates | | | Frequentist Estimates | | |
|--|---|-------|--------|-----------------|-------|--------|-----------------------|-------|--------|
| | AAE | AvCvr | AvLn | AAE | AvCvr | AvLn | AAE | AvCvr | AvLn |
| $x \sim N(15, 5), N(20, 6),$ $N(25, 5), N(28, 8)$ $y \sim 5x + N(0, \sigma)$ Strata Sizes = 250 | | | | | | | | | |
| $\sigma = 10$ | 2.428 | 0.972 | 12.020 | 3.480 | 0.960 | 15.256 | 3.416 | 0.952 | 14.804 |
| $\sigma = 20$ | 3.032 | 0.984 | 15.000 | 3.584 | 0.948 | 17.228 | 3.580 | 0.948 | 16.664 |
| $\sigma = 30$ | 3.516 | 0.980 | 18.288 | 4.040 | 0.960 | 19.808 | 4.020 | 0.952 | 19.328 |
| $x \sim \text{Gamma}(2, 1), \text{Gamma}(6, 1),$ $\text{Gamma}(10, 1), \text{Gamma}(14, 1)$ $y \sim x + \text{Gamma}(\eta, 1)$ Strata Sizes = 300 | | | | | | | | | |
| $\eta = 6$ | 0.377 | 0.936 | 1.734 | 0.413 | 0.924 | 1.855 | 0.414 | 0.916 | 1.775 |
| $\eta = 16$ | 0.493 | 0.916 | 2.111 | 0.517 | 0.908 | 2.160 | 0.507 | 0.900 | 2.092 |
| $\eta = 26$ | 0.535 | 0.920 | 2.456 | 0.550 | 0.920 | 2.486 | 0.553 | 0.908 | 2.425 |
| $x \sim 15\text{Beta}(2, 10), 20\text{Beta}(4, 16),$ $25\text{Beta}(7, 25)$ $y \sim x + N(0, \sigma)$ Strata Sizes = 250 | | | | | | | | | |
| $\sigma = 1.0$ | 0.178 | 0.936 | 0.884 | 0.232 | 0.916 | 1.063 | 0.232 | 0.920 | 1.036 |
| $\sigma = 1.5$ | 0.224 | 0.940 | 1.045 | 0.262 | 0.940 | 1.171 | 0.262 | 0.920 | 1.139 |
| $\sigma = 2.0$ | 0.285 | 0.932 | 1.243 | 0.306 | 0.944 | 1.358 | 0.306 | 0.936 | 1.300 |
| $x \sim \text{Bin}(25, .5), \text{Bin}(35, .7),$ $\text{Bin}(40, .5), \text{Bin}(50, .6)$ $y \sim \lfloor x + N(0, \sigma) \rfloor$ Strata Sizes = 250 | | | | | | | | | |
| $\sigma = 5$ | 0.592 | 0.996 | 3.804 | 0.620 | 0.996 | 3.800 | 0.584 | 1.000 | 3.784 |
| $\sigma = 10$ | 1.216 | 0.976 | 5.732 | 1.192 | 0.976 | 5.804 | 1.192 | 0.968 | 5.656 |
| $\sigma = 15$ | 1.472 | 0.956 | 7.396 | 1.464 | 0.952 | 7.352 | 1.484 | 0.956 | 7.276 |
| $x \sim \lfloor \text{Gamma}(3, 1) \rfloor, \lfloor \text{Gamma}(5, 1) \rfloor,$ $\lfloor \text{Gamma}(7, 1) \rfloor$ $y \sim \lfloor x + N(0, 1) \rfloor$ Strata Sizes = 250 | | | | | | | | | |
| $\sigma = 1$ | 0.348 | 0.984 | 1.972 | 0.376 | 0.980 | 2.120 | 0.376 | 0.976 | 2.084 |
| $\sigma = 3$ | 0.364 | 0.984 | 2.712 | 0.408 | 0.988 | 2.768 | 0.408 | 0.984 | 2.672 |
| $\sigma = 5$ | 0.620 | 0.984 | 3.436 | 0.620 | 0.984 | 3.496 | 0.620 | 0.976 | 3.384 |
| $x \sim U(1:20), U(10:30),$ $U(20:40), U(30:55)$ $y \sim \lfloor x + N(0, \sigma) \rfloor$ Strata Sizes = 250 | | | | | | | | | |
| $\sigma = 5$ | 0.952 | 0.980 | 5.108 | 1.016 | 0.980 | 5.264 | 0.972 | 0.972 | 5.184 |
| $\sigma = 10$ | 1.192 | 0.968 | 6.752 | 1.308 | 0.960 | 6.824 | 1.352 | 0.964 | 6.680 |
| $\sigma = 15$ | 1.732 | 0.948 | 8.336 | 1.756 | 0.948 | 8.348 | 1.724 | 0.944 | 8.136 |

Average Absolute Error (AAE) of Point Estimates

Average Coverage (AvCvr) and Average Length (AvLn) of Interval Estimates

tors. The mean absolute deviation of the two estimators are similar and the coverage and lengths of the interval estimates obtained from the two approaches are similar though the interval estimates from the Polya posterior are consistently slightly longer. The estimators based on the Polya posterior estimator which uses the prior information about the median of the auxiliary characteristic demonstrate an improvement in the absolute error of estimation. For these continuous superpopulation models the Polya posterior estimators using the prior information yield interval estimates of similar or lesser length with similar to greater coverage.

For the Polya posterior ignoring the prior information similar results were found for the simulations using superpopulation models comprising discrete characteristics. In these simulations the performance of the interval estimates based on the Polya posterior incorporating the prior information was similar to that of the other approaches. The relative performance of the point estimates based on this posterior distribution was more mixed for these discrete superpopulation models than for the continuous superpopulation models. In general though, this approach yielded point estimates with similar or smaller error compared to the other approaches.

We examined whether the mean of the Polya posteriors for quantiles of stratified populations exhibit the robustness to the loss function observed by Meeden and Vardeman [4] and Nelson and Meeden [6]. Table 2 presents a summary of the performance of the mean of the Polya posteriors in relation to the performance of the median of the posteriors. The mean of the Polya posterior distribution yielded similar to greatly improved performance relative to the median of the posterior distribution under both loss functions for all but one of the classes of superpopulation models. For this group of superpopulation models the mean did not perform better than the median of the posterior or the standard frequentist estimator under the absolute error loss function. For most of the superpopulation models the means of the Polya posterior performed better than the standard frequentist estimator under both loss functions.

8. Final Remarks:

Often, explicit forms for Polya posterior generated distributions for the quantities of interest can not easily be identified and must be estimated through simulation. Here we demonstrate that the standard Polya posterior can be extended readily for application to stratified populations where the form of the resultant distribution for population quantiles is easily identified. Given the complexity of the identified posterior distribution it may be more convenient to approximate the posterior distribution. However, the argument used to identify the forms of the Polya posterior generated quantile distributions is easily adapted to incorporate a broad array of prior information about an auxiliary characteristic in the formulation of the Polya posterior generated quantile distributions. When incorporating prior information that the median of an auxiliary characteristic falls in a specified interval the resultant Polya posterior median estimators tend to possess attractive frequentist properties. This

Table 2: Summary of Simulation Results: Relative Performance of Mean and Median of Polya Posterior for Point Estimation of Population Median

| Population | Polya Estimates using Prior Information | | Polya Estimates | | Frequentist Estimates | | |
|--|---|-------|-----------------|-------|-----------------------|-------|-------|
| | AAE | RMSE | AAE | RMSE | AAE | RMSE | |
| $x \sim N(15, 5), N(20, 6),$ $N(25, 5), N(28, 8)$ $y \sim 5x + N(0, 10)$ | Posterior Median | 2.428 | 3.090 | 3.480 | 4.326 | 3.416 | 4.210 |
| | Posterior Mean | 2.185 | 2.737 | 3.199 | 3.949 | | |
| $y \sim 5x + N(0, 20)$ | Posterior Median | 3.032 | 3.726 | 3.584 | 4.550 | 3.580 | 4.540 |
| | Posterior Mean | 2.714 | 3.373 | 3.274 | 4.181 | | |
| $x \sim \text{Gamma}(2, 1), \text{Gamma}(6, 1),$ $\text{Gamma}(10, 1), \text{Gamma}(14, 1)$ $y \sim x + \text{Gamma}(6)$ | Posterior Median | 0.377 | 0.475 | 0.413 | 0.519 | 0.414 | 0.515 |
| | Posterior Mean | 0.346 | 0.439 | 0.383 | 0.479 | | |
| $y \sim x + \text{Gamma}(16)$ | Posterior Median | 0.493 | 0.644 | 0.517 | 0.671 | 0.507 | 0.656 |
| | Posterior Mean | 0.455 | 0.589 | 0.470 | 0.605 | | |
| $x \sim 15\text{Beta}(2, 10), 20\text{Beta}(4, 16),$ $25\text{Beta}(7, 25)$ $y \sim x + N(0, 1.0)$ | Posterior Median | 0.178 | 0.226 | 0.232 | 0.296 | 0.232 | 0.296 |
| | Posterior Mean | 0.171 | 0.211 | 0.223 | 0.281 | | |
| $y \sim x + N(0, 1.5)$ | Posterior Median | 0.225 | 0.279 | 0.262 | 0.322 | 0.262 | 0.322 |
| | Posterior Mean | 0.212 | 0.264 | 0.251 | 0.306 | | |
| $x \sim \text{Bin}(25, .5), \text{Bin}(35, .7),$ $\text{Bin}(40, .5), \text{Bin}(50, .6)$ $y \sim \lfloor x + N(0, 5) \rfloor$ | Posterior Median | 0.592 | 0.885 | 0.620 | 0.919 | 0.584 | 0.863 |
| | Posterior Mean | 0.618 | 0.794 | 0.619 | 0.801 | | |
| $y \sim \lfloor x + N(0, 10) \rfloor$ | Posterior Median | 1.216 | 1.572 | 1.192 | 1.565 | 1.192 | 1.544 |
| | Posterior Mean | 1.137 | 1.428 | 1.137 | 1.426 | | |
| $x \sim \lfloor \text{Gamma}(3, 1) \rfloor, \lfloor \text{Gamma}(5, 1) \rfloor,$ $\lfloor \text{Gamma}(7, 1) \rfloor$ $y \sim \lfloor x + N(0, 1) \rfloor$ | Posterior Median | 0.348 | 0.590 | 0.376 | 0.613 | 0.376 | 0.613 |
| | Posterior Mean | 0.383 | 0.480 | 0.414 | 0.511 | | |
| $y \sim \lfloor x + N(0, 3) \rfloor$ | Posterior Median | 0.364 | 0.616 | 0.408 | 0.669 | 0.408 | 0.669 |
| | Posterior Mean | 0.420 | 0.540 | 0.450 | 0.572 | | |
| $x \sim U(1:20), U(10:30),$ $U(20:40), U(30:55)$ $y \sim \lfloor x + N(0, 5) \rfloor$ | Posterior Median | 0.952 | 1.236 | 1.016 | 1.327 | 0.972 | 1.298 |
| | Posterior Mean | 0.896 | 1.118 | 0.940 | 1.184 | | |
| $y \sim \lfloor x + N(0, 10) \rfloor$ | Posterior Median | 1.192 | 1.536 | 1.308 | 1.658 | 1.352 | 1.695 |
| | Posterior Mean | 1.174 | 1.454 | 1.206 | 1.498 | | |

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

type of prior information is not easily exploited using standard frequentist methods. The resultant Polya posterior estimates tend to offer similar to greatly improved performance relative to standard frequentist approaches which do not incorporate this prior information.

References

- [1] William Feller. *An Introduction to Probability Theory and Its Applications, Volume I*. Wiley, New York, 1968.
- [2] Malay Ghosh and Glen Meeden. *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, London, 1997.
- [3] Albert Lo. A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16:1684–1695, 1988.
- [4] Glen Meeden and Stephen Vardeman. A noninformative Bayesian approach to interval estimation in finite population sampling. *Journal of the American Statistical Association*, 86:972–980, 1991.
- [5] David Nelson and Meeden Glen. Using prior information about population quantiles in finite population sampling. *Sankhya*, 60:426–445, 1998.
- [6] David Nelson and Meeden Glen. Noninformative nonparametric quantile estimation for simple random samples. *Journal of Statistical Planning and Inference*, In Press.
- [7] Donald Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9:130–134, 1981.
- [8] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer, New York, 1992.
- [9] R. S. Woodruff. Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47:635–646, 1952.