A noninformative Bayesian approach for two-stage cluster sampling

Glen Meeden*
School of Statistics
University of Minnesota
Minneapolis, MN 55455

March 1998
Revised October 1998

*Research supported in part by NSF Grant DMS 9401191

A noninformative Bayesian approach for two-stage cluster sampling

SUMMARY

The polya posterior gives a noninformative Bayesian justification for various single-stage sampling procedures. Here this type of reasoning is extended to two-stage cluster sampling problems. The frequency properties of the resulting procedures are studied.

Key Words: cluster sampling, Polya posterior, finite population sampling, and noninformative Bayes.

1 Introduction

The Polya posterior is a noninformative Bayesian procedure which can be used when little or no prior information is available. The Polya posterior is related to the Bayesian bootstrap of Rubin (1981). See also Lo (1988). One advantage of the Polya posterior is that it has a stepwise Bayes justification and leads to admissible procedures. For further discussion of the Polya posterior see Ghosh and Meeden (1997). In this note we will show how the Polya posterior can be adapted to two-stage cluster sampling. We will consider the simple case where there are N clusters or primary sampling units. Each cluster consists of M subunits or second stage sampling units. We assume that first a random sample without replacement of n clusters are selected from the N clusters. Then within each selected cluster a random sample without replacement of m subunits is chosen. In section two the two-stage Polya posterior for this problem will be described. In section three an asymptotic expression for the variance of the population mean of the subunits given the observed sample under the two-stage Polya posterior is found. This formula is very similar to the usual frequentist estimate of variance for this problem. In section four the two-stage Polya posterior is shown to have a step-wise Bayesian justification and an admissibility result is proved. These facts indicate that the two-stage Polya posterior will yield both point and interval estimators for a variety of population quantities that should have good frequentist quantities. In section five some concluding remarks are given.

2 A two-stage Polya posterior

We begin by recalling some facts about the Polya posterior for the most basic situation, simple random sampling without replacement. Given the data the 'Polya posterior' is a predictive joint distribution for the unobserved or unseen units in the population conditioned on the values in the sample. Given a sample we construct this distribution as follows. We consider an urn that contains the same number of balls as there are units in the sample and where the value of each observed unit in the sample is assigned to a unique ball in the urn. The distribution is specified as follows. We begin by choosing a ball at random from the urn and assigning its value to the unobserved unit in the population with the smallest label. This ball and an additional ball with the same value are returned to the urn. Then another ball is chosen at random from the urn and its value is assigned to the unobserved unit in the population with the second smallest label. This second ball and another with the same value are returned to the urn. This process is continued until all the unobserved units in the population are assigned a value. Once they have all been assigned values we have observed one realization from the 'Polya posterior'. Hence by simple Polya sampling we have a predictive distribution for the unseen given the seen. A good reference for Polya sampling is Feller (1968).

For two stage cluster sampling we can modify the above by using Polya sampling for both stages. In what follows we will always assume that for any pair of clusters the distinct values that appear in one cluster are either identical to the distinct values of the other cluster or the two sets of distinct values contain no common member. For simplicity assume that all the observed values are distinct. Consider a sample which consists of n clusters where within each of them we have observed m subunits and by assumption all values of the subunits are distinct. Then these n sets of distinct values are placed into n envelopes. These n envelopes are then placed into a large urn and Polya sampling from this large urn, where the envelopes are the basic elements sampled, is used to distribute the sets of distinct observed cluster values to the unobserved clusters. Once this is done each cluster has associated with it a set of possible values. For the clusters in the sample these are just the values of the subunits in the sample. For the unobserved clusters these are the values in the envelope assigned to it under the first stage of Polya sampling. To complete a simulated copy of the entire population we just do Polya sampling within each cluster, using the observed or assigned values for each cluster, independently across the clusters until each cluster contains M values. Clearly this defines a predictive distribution for the unseen given the seen when two-stage cluster sampling is used. The labels within clusters really play no role in this process.

It has been shown that for a variety of decision problems, procedures based on the 'Polya posterior' are generally admissible because they are stepwise Bayes. One of the goals of this note is to show that this is the case for the two-stage Polya posterior outlined just above. In these earlier stepwise Bayes arguments a finite sequence of disjoint subsets of the parameter space is selected, where the order is important. A different prior distribution is defined on each of the subsets. Then the Bayes procedure is found for each sample point that receives positive probability under the first prior. Next the Bayes procedure is found for the second prior for each sample point which receives positive probability under the second prior and which was not taken care of under the first prior. Then the third prior is considered and so on. To prove the admissibility of a given procedure one must select the sequence of subsets, their order, and the sequence of priors appropriately. We will now see how this basic argument can be modified to demonstrate admissibility for the problem at hand.

As noted before in what follows we will always assume that for any pair of clusters the distinct values that appear in one cluster are either identical to the distinct values of the other cluster or the two sets of distinct values contain no common member. In this section we will not formally specify the parameter space nor the actual order of the subsets of the parameter space to be used in the argument. Instead we will assume that a typical subset is at hand and define the corresponding prior distribution. Then we will find the posterior for a typical sample point at this stage. We will delay until section four the demonstration of admissibility.

To this end let y_{ij} denote the value of the characteristic of interest for subunit j in cluster i for $i = 1, \ldots, N$ and $j = 1, \ldots, M$. Let y be the y_{ij}'s in lexicographic order. Let $b^1, b^2, \ldots, b^r$ be r vectors of real numbers where $1 \le r \le N$ and the values of each b^l are distinct and they have no values in common. Let $|b^l|$ denote the length of b^l and let b^l_k be the kth element of b^l. For each b^l we assume that $|b^l| \le M$. A vector b^l denotes a set of possible values that may appear in a cluster at this stage of the argument. Let t be a vector of length N where each t_i may take on the values $1, 2, \ldots, r$. Associated with each vector y there is a vector t. When the subunits in cluster i take on the values in vector b^l this is denoted by $t_i = l$.

The prior distribution for y is defined by first defining a marginal distribution for t and then a conditional distribution for y given t. Before stating the actual definition we need a bit more notation. For a given t, y and $l = 1, \ldots, r$ let

$$c_t(l) = \text{the number of } t_i\text{'s which equal } l$$

and for each $i = 1, \ldots, N$ and $k = 1, \ldots, |b^{t_i}|$ let

$$c^i_{t,y}(k) = \text{the number of } y_{ij}\text{'s which equal } b^{t_i}_k$$

Finally we will be restricting the prior to y's for which $c_t(l) \ge 1$ for all l and

$c^i_{t,y}(k) \geq 1$ for all i and k. On this set the prior is given by

$$\begin{aligned} p(y) &= p(t)p(y|t) \qquad (1) \\ &\propto \int_0^1 \cdots \int_0^1 \prod_{l=1}^{r} \theta_l^{c_t(l)-1} d\theta_1 \cdots d\theta_r \\ &\quad \times \prod_{i=1}^{N} \int_0^1 \cdots \int_0^1 \prod_{k=1}^{|b^{t_i}|} \theta_k^{c^i_{t,y}(k)-1} d\theta_1 \cdots d\theta_{|b^{t_i}|} \\ &\propto \frac{\prod_{l=1}^{r} \Gamma(c_t(l))}{\Gamma(N)} \prod_{i=1}^{N} \frac{\prod_{k=1}^{|b^{t_i}|} \Gamma(c^i_{t,y}(k))}{\Gamma(M)} \end{aligned}$$

where within each Dirichlet integral it is assumed that the θ's sum to one. It should be noted that our assumptions place some restrictions on the samples that can be observed. For example if $M = 3$ and $m = 2$ it would not be possible to observe three clusters with the values $(0.11, 0.22)$, $(0.22, 0.33)$ and $(0.33, 0.44)$ since they could only arise from from a common b vector which contained all four of the distinct values. This cannot happen since the length of b can be at most 3, the value of M.

This prior is closely related to the typical prior used in the stepwise Bayes argument which generates the Polya posterior for the simple case described above. In fact it is just that prior applied to t and then applied again independently within each cluster conditional on the values in t. This makes it easy to find the marginal probability of a sample and the conditional probability of a parameter point given the sample. For more details on the following argument see sections 2.3 and 2.6 of Ghosh and Meeden (1997).

First we need to introduce some notation to denote a possible two-stage sample. We begin by letting h be the vector of cluster indices that appear in the sample arranged in increasing order. So h is of length n. For $i = 1, \ldots, n$ let s_i denote the labels of the subunits which appear in the sample in cluster h_i. Let t_h be t restricted to the clusters appearing in h. Hence t_{h_i} identifies the vector b^l which contains the set of distinct values that appeared in the sampled cluster h_i. Finally let y_{s_i} denote the actual set of m observed values for the sampled subunits in cluster h_i. We will denote the entire sample by $z = (h, t_h, s, y_s)$ where $s = (s_1, \ldots, s_n)$ and $y_s = (y_{s_1}, \ldots, y_{s_n})$.

Let $\{a_1, \ldots, a_w\}$ be the distinct values that appear in y_s. We say that $a_i \sim a_j$ if there exists distinct elements $a_{u_1}, \ldots, a_{u_v}$ such that $a_{u_1} = a_i$, $a_{u_v} = a_j$ and for $k = 1, \ldots, v-1$ a_k and a_{k+1} appear together in some y_{s_q}.

Clearly this defines an equivalence relation over the a_u's. Moreover for a given sample the distinct values making up an equivalence class must be a subset of the values making up b^l.

As we shall see at this stage of the stepwise Bayes argument the only samples which we can observe which have not been accounted for at an earlier stage are those for which the equivalence classes of $\{a_1, \ldots, a_w\}$ are exactly $b^1, b^2, \ldots, b^r$.

For a fixed z and $l = 1, \ldots, r$ we let

$$c_{t_h}(l) = \text{the number of } t_{h_i}\text{'s which equal } l$$

and for each $i = 1, \ldots, n$ and $k = 1, \ldots, |b^{t_{h_i}}|$ let

$$c^{h_i}_{t_{h_i}, y_{s_i}}(k) = \begin{cases} \text{the number of } y_{h_ij}\text{'s in } y_{s_i} \text{ which equal } b_k^{t_{h_i}} & \text{if } b_k^{t_{h_i}} \in y_{s_i} \\ 1 & \text{if } b_k^{t_{h_i}} \notin y_{s_i} \end{cases}$$

Given z we can break t into two parts t_h and $t_{\overline{h}}$ where $\overline{h}$ is the complement of h, i.e. $\overline{h}$ is the collection of clusters not appearing in the sample. Note that $t_{\overline{h}_i}$ identifies the set of distinct values which will appear in the unsampled cluster $\overline{h}_i$. Now $p(y)$ can be written as

$$p(y) = p(t_h, t_{\overline{h}})p(y_h|t_h)p(y_{\overline{h}}|t_{\overline{h}})$$

where y_h are the values of the subunits in the clusters h and $y_{\overline{h}}$ are the values of the subunits in the clusters $\overline{h}$. To find $p(z)$ we just need to sum the above over all y which are consistent with the observed sample z. The $p(y_{\overline{h}}|t_{\overline{h}})$ must sum to one since for each fixed $t_{\overline{h}}$ there are no constraints on the values in $y_{\overline{h}}$. Then using the standard argument we find that

$$p(z) = \frac{\prod_{l=1}^{r} \Gamma(c_{t_h}(l))}{\Gamma(n)} \prod_{i=1}^{n} \frac{\prod_{k=1}^{|b^{t_{h_i}}|} \Gamma(c^{h_i}_{t_{h_i}, y_{s_i}}(k))}{\Gamma(m)}$$

Now let t' and y' denote a parameter point which is consistent with the sample point z. Then following the argument for the usual Polya posterior we have

$$
\begin{aligned}
p(y'|z) &= p(y')/p(z) \qquad (2)\\
&= \Big\{\prod_{l=1}^{r}\Big\{\Gamma(c_{t'}(l))/\Gamma(c_{t_h}(l))\Big\}\Big\}/\Big\{\Gamma(N)/\Gamma(n)\Big\}\\
&\quad\times\prod_{i=1}^{n}\Big\{\Big\{\prod_{k=1}^{|b^{t'}_{h_i}|}\Big\{\Gamma(c^{h_i}_{t',y'}(k))/\Gamma(c^{h_i}_{t_{h_i},y_{s_i}}(k))\Big\}\Big\}/\Big\{\Gamma(M)/\Gamma(m)\Big\}\Big\}\\
&\quad\times\prod_{i=1}^{N-n}\Big\{\Big\{\prod_{k=1}^{|b^{t'}_{\bar{h}_i}|}\Gamma(c^{\bar{h}_i}_{t',y'}(k)\Big\}/\Big\{\Gamma(M)/\Gamma(|b^{t'}_{\bar{h}_i}|)\Big\}\Big\}
\end{aligned}
$$

It is easy to see that equation 2 is indeed the mathematical formulation of the two-stage Polya posterior described early in this section when all the values of all the observed subunits are distinct. More generally the first main factor represents the probability structure under Polya sampling for assigning observed sets of distinct cluster values to the unsampled clusters. The second main factor represents Polya sampling within the sampled clusters to simulate completed clusters. The last main factor represents Polya sampling within the unsampled clusters to simulate completed clusters using the assigned set of distinct values.

3 Asymptotic variance

In many situations in cluster sampling the y_{ij}'s are all distinct. Although the formulation in the previous section is not completely general it does cover this setup. In this special case we will find the asymptotic variance of the population mean of the subunits given the observed sample z under the distribution given in equation 2. To do this we need to introduce some additional notation. For a given y we let

$$\mu_i = \mu_i(y) = \sum_{j=1}^{M} y_{ij}/M$$

and

$$\mu = \mu(y) = \sum_{i=1}^{N} \mu_i/N$$

be the population mean of the subunits. Let $f_1 = n/N$ and $f_2 = m/M$ be the respective sampling proportions. For a given sample $z = (h, t_h, s, y_s)$ we let

$$\overline{y}_{s_i} = \sum_{j \in s_i} y_{ij}/m,$$

$$\overline{\overline{y}}_s = \sum_{i=1}^{n} \overline{y}_{s_i}/n,$$

$$\mathrm{Var}_1(y_s) = \sum_{i=1}^{n} (\overline{y}_{s_i} - \overline{\overline{y}}_s)^2/(n-1),$$

$$\mathrm{Var}_{2,i}(y_{s_i}) = \sum_{j \in s_i} (y_{ij} - \overline{y}_{s_i})^2/(m-1)$$

and

$$\mathrm{Var}_2(y_s) = \sum_{i=1}^{n} \mathrm{Var}_{2,i}(z)/n$$

If the sampling design at each stage is just simple random sampling without replacement then an unbiased estimator of the variance of $\overline{\overline{y}}_s$ is just

$$\frac{1-f_1}{n} \mathrm{Var}_1(y_s) + \frac{f_1(1-f_2)}{mn} \mathrm{Var}_2(y_s) = V_1 + V_2 \tag{3}$$

say. (See for example Cochran (1977).) We can now state and prove a theorem about the variance of $\overline{\overline{y}}_s$ under the two stage Polya posterior for a given sample z.

Theorem 1. *Given a sample $z = (h, t_h, s, y_s)$ where all the values in y_s are distinct, N is large and n/N is small then under the probability distribution given in equation 2 an approximate expression for the conditional variance of the population mean given z is*

$$\mathrm{Var}(\mu|z) \doteq V_1 \frac{n-1}{n+1} \frac{N}{N-n} + V_2 \frac{m-1}{m+1}$$

where V_1 and V_2 are as in equation 3.

Proof. Since all the values in y_s are distinct the distribution in equation 2 is based upon n vectors $b^1, \dots, b^n$ where b^l is just the values making up y_{s_l}. Now given z the first main factor in the posterior in equation 2 represents Polya sampling from the set of n observed cluster values to the $N-n$ unsampled clusters. Let $t^*_{\overline{h}}$ be one fixed possible set of outcomes for this Polya sampling. Let $t^* = (t_h, t^*_{\overline{h}})$ then using the well known formula that the variance is equal to the expectation of the conditional variance plus the variance of the conditional expectation we have that

$$V(\mu|z) = E(V(\mu|z,t^*)) + V(E(\mu|z,t^*)) \tag{4}$$

where z is fixed throughout the proof.

Now $c_{t^*}(l)$ is just the number of clusters were the values y_{s_l} will appear when constructing a simulated copy of the entire population conditioned on t^*. Each such cluster will be completed using Polya sampling. If j denotes one such cluster then we have

$$E(\mu_j|z,t^*) = \overline{y}_{s_l}$$

and by page 46 of Ghosh and Meeden (1997)

$$V(\mu_j|z,t^*) = \frac{1-f_2}{m}\,\mathrm{Var}_{2,l}(y_{s_l})\frac{m-1}{m+1}$$

Finally it is well known that when N is large and n/N is small that

$$\left(\frac{c_{t^*}(1)}{N}, \dots, \frac{c_{t^*}(n)}{N}\right) \sim \text{Dirichlet}(1, \dots ,1)$$

We first evaluate the second term in the right hand side of equation 4. Since

$$\begin{aligned} E(\mu|z,t^*) &= N^{-1}\sum_{i=1}^{N} E(\mu_i|z,t^*) \\ &= \sum_{l=1}^{n} \overline{y}_{s_l}\frac{c_{t^*}(l)}{N} \end{aligned} \tag{5}$$

we have

$$
\begin{aligned}
V(E(\mu|z,t^*)) &= V\Big(\sum_{l=1}^{n} \overline{y}_{s_l} \frac{c_{t^*}(l)}{N}\Big) \\
&= \sum_{l=1}^{n} \overline{\overline{y}}_{s_l}^2 V\Big(\frac{c_{t^*}(l)}{N}\Big) + 2\sum_{l<k} \overline{y}_{s_l}\overline{y}_{s_k} Cov\Big(\frac{c_{t^*}(l)}{N}, \frac{c_{t^*}(k)}{N}\Big) \\
&= \frac{n-1}{n^2(n+1)} \sum_{l=1}^{n} \overline{\overline{y}}_{s_l}^2 - \frac{2}{n^2(n+1)} \sum_{l<k} \overline{y}_{s_l}\overline{y}_{s_k} \\
&= \frac{1}{n(n+1)} \sum_{l=1}^{n} (\overline{y}_{s_l} - \overline{\overline{y}}_s)^2 \\
&= V_1 \frac{n-1}{n+1} \frac{N}{N-n} \qquad (6)
\end{aligned}
$$

We now evaluate the first term in the right hand side of equation 4. Since

$$
\begin{aligned}
V(\mu|z,t^*) &= N^{-2} \sum_{i=1}^{N} V(\mu_i|z,t^*) \\
&= N^{-2} \sum_{l=1}^{n} V(\mu_{h_l}|z,t^*) c_{t^*}(l) \\
&= \frac{1}{N} \frac{1-f_2}{m} \frac{m-1}{m+1} \sum_{l=1}^{n} \mathrm{Var}_{2,l}(y_{s_l}) \frac{c_{t^*}(l)}{N}
\end{aligned}
$$

we have

$$
\begin{aligned}
E(V(\mu|z,t^*)) &= \frac{1}{N} \frac{1-f_2}{m} \frac{m-1}{m+1} \sum_{l=1}^{n} \mathrm{Var}_{2,l}(y_{s_l}) E\Big(\frac{c_{t^*}(l)}{N}\Big) \\
&= \frac{1}{N} \frac{1-f_2}{m} \frac{m-1}{m+1} \mathrm{Var}_2(y_s) \\
&= V_2 \frac{m-1}{m+1} \qquad (7)
\end{aligned}
$$

Substituting equations 6 and 7 into equation 4 we see that the proof is complete. □

Under the assumptions of the theorem and if m is moderately large, say $m > 20$, we see that the approximation for $\mathrm{Var}(\mu|z)$ is very close to the variance of $\overline{\overline{y}}_s$ under two-stage simple random sampling given in equation 3.

It is easy to find the value of the point estimate of μ under the posterior given in equation 2 with squared error loss when all the observed values in y_s are distinct. Using equation 5 we see that

$$\begin{aligned} E(\mu|z) &= E(E(\mu|z,t^*)) \\ &= \sum_{l=1}^{n} \overline{y}_{s_l} E\left(\frac{c_{t^*}(l)}{N}\right) \\ &= \overline{\overline{y}}_s \end{aligned} \tag{8}$$

which is the usual point estimate of μ.

In the next section we will prove an admissibility result for the family of posteriors given in equation 2.

4 An admissibility result

In this section we will demonstrate an admissibility result under two stage random sampling. Actually the design plays no role in the proof and is assumed just for convenience. For further details see Scott (1975).

Let $b = (b_1, \ldots, b_K)$ be a vector of K distinct real numbers. We assume that $K \geq 2$ and in fact it may be rather large. The parameter space used in the proof will depend on the vector b. For a positive integer $r \geq 1$ let $b^1, \ldots, b^r$ denote r vectors where the length of each vector is less than or equal to M. The values of each b^l for $l = 1, \ldots, r$ are all distinct and belong to the vector b. Furthermore the b^l's contain no common values. For such a fixed collection, $b^1, \ldots, b^r$, we let

$$\begin{aligned} \mathcal{Y}(b^1, \ldots, b^r) = \{(t, y) \ : \ & \text{such that for } i = 1, \ldots, N, \ \ t_i = l \text{ for some} \\ & l = 1, \ldots, r \text{ and each } b^l \text{ appears at least once}\} \end{aligned}$$

Note that $\mathcal{Y}(b^1, \ldots, b^r)$ corresponds to all those populations where the distinct values appearing in each cluster is one of the b^l's and each b^l must appear at least once. Also for a given r the number of possible different sets of $b^1, \ldots, b^r$ that can be constructed depends on K and M. Moreover for r sufficiently large no such possible sets exist. We let $\mathcal{Y}(b_1, \ldots, b_K)$ denote the parameter space which consists of all populations which belong to some $\mathcal{Y}(b^1, \ldots, b^r)$ for some $r = 1, 2, \ldots$. This parameter space consists of all those populations where the set of distinct values appearing in a cluster is a subset of b and the set of distinct values appearing in two cluster are either disjoint or identical.

Theorem 2. *Assume two-stage simple random sampling without replacement. For estimating a real valued function $\gamma(y)$ with a strictly convex loss function when the parameter space is $\mathcal{Y}(b_1, \dots, b_K)$ the posterior in equation 2 has a stepwise Bayes justification and hence will yield an admissible estimator for $\gamma(y)$.*

Proof. Because of discussion earlier in the paper the part of the proof which remains to be considered is the selection of the subsets of the parameter space and their proper order in the stepwise Bayes argument. The argument will consist of several main stages.

In the first main stage we will consider $b^1, \dots, b^r$'s were each b^l contains just one element. This stage will have several smaller steps. In the first step we let $r = 1$ and get all those parameter points where every subunit takes on the same value. There will be K different cases to consider which can be taken in any order. There is one possible parameter point for each case which yields one possible sample point. Both the prior and posterior put mass one on the single parameter point which agrees with equation 2 in this case.

In the next step we let $r = 2$ and we have $\binom{K}{2}$ cases which can be handled in any order. For example when $b^1 = (b_1)$ and $b^2 = (b_2)$ the prior in equation 1 depends only on $p(t)$ and the only sample points which will appear which were not taken care of in the previous step are those were each sampled cluster is either all b_1's or all b_2's and each type appears at least once. It is easy to check that for such sample points the prior in equation 1 yields the posterior in equation 2.

The next step where $r = 3$ is handled in the same manner. We continue in this way until $r = \text{minimum}\{M, K\}$ which completes the first main stage.

The second main stage will consider $b^1, \dots, b^r$'s were each b^l contains either one or two elements. Again this stage will have several smaller steps. In the first step we will consider the cases where only one b^l has two distinct values. This is broken up into several cases as well. In the first $r = 1$ and b^1 contains just two elements. There are $\binom{K}{2}$ such cases which can be handled in any order so assume that $b^1 = (b_1, b_2)$ and consider the prior which puts mass one on the parameter point were the unique values in each cluster are just b_1 and b_2. Under this prior each observed cluster will either contain just b_1 or just b_2 or both. The only samples which have not been taken care of previously are those where for at least one sampled cluster both b_1 and

b_2 appear. For such samples the prior in equation 1 yields the posterior in equation 2. In the next step we consider the cases where one b^1 contains two elements and the remaining b^l's contain just one element where by necessity all the elements that appear are distinct. Again we have $\binom{K}{2}$ choices for the two values which appear in b^1. These can be handled in any order. For definiteness assume that $b^1 = (b_1, b_2)$ and there are some additional b^l's which contain just one element. Then the samples which are possible now and have not been considered earlier are those where for at least one sampled cluster the set of unique observed values are b^1 and for the rest of the sampled clusters just one valued was observed where at least one these values is other than b_1 or b_2. The order in which these cases are considered depends just on the b^l's for $l \neq 1$ and just follows the order in the first main stage.

In the next step we begin with $r = 2$ where b^1 and b^2 each contain two elements and these can be handled in any order. Then we consider cases where b^1 and b^2 both contain two elements and the remaining b^l's contain just one element. Again for fixed choices of b_1 and b_2 the order depends only on the b^l's for $l > 2$ and just follows the order in the first main stage.

In the next step we handle the cases where exactly three of the b^l's contain just two elements and the rest only one. Then the next step takes care of the cases where exactly four of the b^l's contain just two elements and the rest only one. We continue until all possible samples in which observed clusters contain either one or two distinct values have been handled. It is easy to check in all cases that for the sample points being considered the prior in equation 1 yields the posterior in equation 2. This completes the second main stage of the argument.

In the third main stage of the argument we handle all those sample points where at least one of the observed equivalence classes contains three distinct values and none of them contain more than three. The first step considers cases were $r = 1$ and b^1 contains three elements. Then we consider all those cases where just one b^l contains three distinct values and the rest either one or two distinct values. The order follows that of the first two main stages. Then we handle the cases where exactly two of the b^l's have three elements, then where exactly three of them have three elements, then four and so on until all possible sample points have been taken care of.

The next main stage handles all those sample points where at least one of the observed equivalence classes contains four distinct values and none of them contain more than four. This is followed by the case for five values and so on. The number of main stages and the number of cases within a main stage depend on the sizes of N, M and K. As before it is easy to check in all cases that for the sample points being considered the prior in equation 1 yields the posterior in equation 2. This concludes the proof. $\square$

5 Concluding remarks

It might seem from equation 8 that Theorem 2 proves the admissibility of the estimator $\overline{\overline{y}}_s$ when estimating the population mean with squared error loss. In fact the theorem holds for a slightly different estimator which agrees with $\overline{\overline{y}}_s$ only when all the observations in the sample are distinct. To see why this is so note the following. Suppose we have a sample where the distinct values in one of the observed clusters are 0.11 and 0.22 while in a second they are 0.11, 0.22 and 0.33 and in a third they are 0.11, and 0.44 with different values appearing in the other sampled clusters. Note these four values belong to the same equivalence class and the posterior in equation 2 assumes that all four values must belong to each of the three sampled clusters since we assumed that the distinct values in clusters are either disjoint or identical. In many situations such an assumption is not unreasonable and it does seem to be in the spirit of a noninformative Bayesian approach. However other assumptions could also be sensible. The posterior in equation 2 depends on the definition of $\mathcal{Y}$ which in turn depends on the assumption that the distinct values appearing in any two clusters are either identical or distinct. This assumption allows one to identify unambiguously the set of distinct values associated with each sampled cluster. Without this assumption then it becomes possible that certain sampled clusters may have arisen from a whole collection of possible 'parent' clusters. For example for the above three sampled clusters how should one take into account that the first cluster could contain either 0.33 or 0.44 or perhaps neither of them? It is not clear what a sensible noninformative Bayesian choice for a collection of parent clusters should be for this example. Moreover both the stepwise Bayes argument and the problem of simulating from the posterior become much more complicated for problems where more than one parent is allowed. Nevertheless it would be of interest to prove a theorem without this assumption and extend it to the unbalanced case where the cluster size and sample size for a cluster may vary with the cluster.

One of the advantages of the Polya posterior is that it often yields Bayesian credible intervals with good frequentist properties. We see from Theorem 1 and equation 3 that this should be true when m is large enough. If m gets too small however the 0.95 credible intervals become to short and may cover significantly less that 95% of the time. However when $m \geq 20$ the two-stage Polya posterior should yield good frequentist intervals not only for the population mean but also other functions of interest of the population.

References

[1] William G. Cochran. *Sampling Techniques.* Wiley, New York, 1977.

[2] William Feller. *An introduction of probability theory and its applications, volume I.* Wiley, New York, 1968.

[3] Malay Ghosh and Glen Meeden. *Bayesian Methods for Finite Population Sampling.* Chapman and Hall, London, 1997.

[4] Albert Lo. A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16:1684–1695, 1988.

[5] Donald Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9:130–134, 1981.

[6] Alastair Scott. On admissibility and uniform admissibility in finite population sampling. *Annals of Statistics*, 3:489–491, 1975.