Journal of Survey Statistics and Methodology (2014) 2, 38–57

MORE EFFICIENT INFERENCES USING RANKING INFORMATION OBTAINED FROM JUDGMENT SAMPLING

GLEN MEEDEN*
BO RA LEE

Judgment poststratification is an extension of ranked set sampling. It arises when sampled units from a population can be ranked rather easily without actually measuring the variable of interest and when determining the actual values can be expensive and time consuming. Under such a scheme, samples will contain units for which the variable of interest is observed while for others only that they are larger or smaller than one of the observed units. This paper will argue that standard methods ignore information contained in such samples and show that an objective stepwise Bayes analysis based on the Polya posterior leads to improved inferential procedures.

1. INTRODUCTION

When seeking to effectively estimate the yield of a pasture, McIntyre (1952) proposed a sampling method that later became known as ranked set sampling (RSS). This method assumes that it is possible to efficiently and cheaply compare two units in the population, but determining the actual value of a unit is much more difficult. In addition, there are fairly strict conditions on the relationship within the sample between the units that are fully observed and those that only have ranking information. Despite this restriction, in practice, there are many situations in which this approach can be applied. A helpful recent summary of the literature is given by Chen, Bai, and Sinha (2004).

Recently, MacEachern, Stasny, and Wolfe (2004) introduced judgment poststratification (JPS), which is a less restricted form of RSS. In JPS, a simple

Glen Meeden is a Professor of Statistics in the School of Statistics, University of Minnesota, Minneapolis, MN 55455. Bo Ra Lee was a graduate student in the School of Statistics, University of Minnesota, Minneapolis, MN 55455, at the time this research was performed.
*Address correspondence to Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455; E-mail: gmeeden@umn.edu

doi: 10.1093/jssam/smt020 Advance access publication 2 February 2014

random sample is again combined with ranking information based on the judgment of an expert. First, one draws a simple random sample of size n and observes the variable of interest. Next, for some $m > 1$ and for each unit in the original sample, an additional random sample of size $m - 1$ is taken. An expert orders the set containing the observed unit, along with the additional $m - 1$ units, and the rank of the observed unit within this set is noted. This is done for each unit in the sample. When this process is completed, the data associated with the ith sampled unit are its observed value and rank, (y_i, r_i). These rankings are used to poststratify the sample of fully observed units, which yields an estimator for the population total or mean. Frey and Feeman (2012) showed that this JPS estimator is inadmissible within a certain class and derived alternative estimators that are admissible. Although JPS tends to be less efficient than the more balanced RSS, it is more flexible and allows for ties.

In the Bayesian approach to survey sampling, information about the population is incorporated into a prior distribution. After the sample is observed, inferences are based on the posterior distribution of the unobserved units in the population, given the values of the observed units in the sample. Bayes methods have infrequently been used in practice because it is difficult to find sensible prior distributions.

In the stepwise Bayes approach, given a sample, inference is still based on a posterior distribution; however, the collection (for all possible samples) of the posteriors does not arise from a single prior, but from a whole family of prior distributions. Many of the standard estimators can be given a stepwise Bayesian interpretation. For a situation in which one believes that the observed units are roughly exchangeable with the unobserved units, the appropriate stepwise Bayes posterior distribution is the Polya posterior. By using this approach, one can prove the admissibility of the sample mean as an estimator for the population mean. Details can be found in the work of Ghosh and Meeden (1997).

Lazar, Meeden, and Nelson (2008) demonstrated that the Polya posterior can be constrained to account for certain kinds of prior information, such as the population means of auxiliary variables. This paper will argue that the usual RSS and JPS estimators do not make use of all of the information in the sample. We will show that an appropriately constrained version of the Polya posterior yields admissible estimators that are better than the RSS and JPS estimators. This is because the Polya posterior, in a natural way, more fully exploits the information contained in the ranks.

In Section 2, we describe in more detail the estimators proposed by Frey and Feeman (2012). In Section 3, we briefly review the Polya posterior and show how it can be constrained to use all of the information contained in a JPS. Next, through some simulations, we compare our estimators to standard methods. We explain the reasons for the superior performance of our estimator. In Section 4, we more formally discuss the theory underlying our estimator

and explain how it is computed in practice. In Section 5, we show how our approach can yield approximate 95% confidence intervals for population parameters. In Section 6, we present another simulation study using the selling prices of a population of recently sold houses as the population of interest. Section 7 contains some final remarks.

2. JUDGMENT POSTSTRATIFICATION ESTIMATORS

Consider a finite population of size N, where $y = (y_1,\ldots,y_N)$ is the unknown characteristic of interest. Let s denote a simple random sample of size n and $y(s) = \{y_i\colon i \in s\}$ the observed values. MacEachern, Stasny, and Wolfe (2004) considered the following situation. Let $m > 1$ be given, then for each $i \in s$, an additional $m-1$ units are selected at random from the population. Without actually observing their y values, an expert ranks these $m-1$ units, along with the observed y_i. This results in a rank for y_i, r_i. Here, r_i must represent a value in the set $\{1,2,\ldots,m\}$. Once this has been done for each $i \in s$, we can use the set of ranks, $\{r_i\colon i \in s\}$, to poststratify the sample. That is, for $j = 1,\ldots,m$, the jth poststratum consists of all of those units in the sample whose rank is equal to j. It is possible for a stratum to be empty. For $j = 1,\ldots,m$, let $\bar{Y}_{(j)}$ be the mean of the units in the sample whose rank is j. When all of the poststrata contain at least one unit, the JPS estimator for the population mean is just the average of the means of the poststrata. When there are empty poststrata, the standard JPS mean estimator is the average of the sample means for the nonempty poststrata.

Frey and Feeman (2012) proved that this estimator is inadmissible under squared error loss within a certain class of linear estimators. In addition, they derived alternate estimators that are admissible in this class and showed that one of them is always better than the standard estimator. This estimator is given by

$$\delta_{\mathrm{JPS}} = \sum_{j=1}^{m} w_j \bar{Y}_{(j)} \tag{1}$$

where

$$w_j = (n_j/(mn_j + 2))/\sum_{k=1}^{m}(n_k/(mn_k + 2)) \tag{2}$$

where n_j is the number of units in the sample that belong to the jth poststratum. The estimator is well defined when $n_j = 0$ if we set $\bar{Y}_{(j)} = 0$.

In the next section, we will describe a new way to use the information contained in those units in the sample that were used to produce the rank for a

fully observed unit in the sample. However, first, we review some facts about the Polya posterior.

3. USING MORE OF THE INFORMATION IN THE SAMPLE

3.1 The Polya Posterior

We begin by introducing some terminology. We will call an observation in the original sample for which we observe the y value an "elder." An observation in the subsequent samples for which we only have order information will be called a "sib." For the moment, assume that there are no sibs in the sample; that is, we are in the standard setup. In this case, the Polya posterior is based upon Polya sampling from an urn. It works as follows: suppose that the values from n observed units are marked on n balls and placed in urn I. The remaining unseen $N-n$ units of the population are represented by $N-n$ unmarked balls placed in urn II. One ball from each urn is drawn with equal probability and the ball from urn II is assigned the value of the ball from urn I. Both balls are returned to urn I. Thus, at the second stage of Polya sampling, urn I has $n+1$ balls and urn II has $N-n-1$ balls. This procedure is repeated until urn II is empty, at which point the N balls in urn I constitute one complete simulated copy of the population. Any finite population quantity—means, totals, or regression coefficients—may now be calculated from the complete copy. By creating K complete copies in the same manner, the Polya posterior is generated for the desired population quantity. The mean of these K simulated values is the point estimate.

For $i \in s$, let p_i denote the proportion of units in a full, simulated copy of the population that have the value y_i. One can show that under the Polya posterior, $E(p_i) = 1/n$; from this, it follows that under the Polya posterior, the posterior expectation of the population mean is the sample mean. Clearly, under this scheme, the simulated values for the unseen are correlated and one can show that the posterior variance of the population mean is $(n-1)/(n+1)$ times the usual design-based variance of the sample mean under simple random sampling without replacement.

The Polya posterior has a decision-theoretic justification based on its stepwise Bayesian nature. Using this fact, many standard estimators can be shown to be admissible. Details can be found in the work of Ghosh and Meeden (1997). The Polya posterior is the Bayesian bootstrap of Rubin (1981) applied to finite population sampling. Lo (1988) also discussed the Bayesian bootstrap in finite population sampling. Some early related work can be found in Hartley and Rao (1968) and Binder (1982).

Although both the Polya posterior and the usual bootstrap methods in finite population sampling are based on the notion of exchangeability, the logic that underlies these methods is different. Gross (1980) introduced the

basic idea for the bootstrap. Assume simple random sampling without replacement and suppose that $N/n = m$ is an integer. We create a reasonable guess for the population by combining m replicates of the sample. By taking repeated random samples of size n from this created population, we can study the behavior of an estimator of interest. Booth, Butler, and Hall (1994) studied the asymptotic properties of such estimators. This is in contrast to the Polya posterior, which considers the sample to be fixed and repeatedly generates complete versions of the population. This, in turn, generates a distribution for the population parameter of interest. Inferences for the population parameter are made by using this predictive distribution. Both approaches, however, assume that the unseen units in the population can only have values that have appeared in the sample. This is obviously a fiction, but in both cases, it is a close enough approximation to reality to yield good estimators when the sampling design is simple random sampling without replacement.

3.2 The Polya Posterior and Judgment Sampling

In judgment sampling, it is the rank of an elder within a set of sibs that is important. This focus on the ranks in the standard approach results in the requirement that every elder must have the same number of sibs. Rather than just using the ranks of the elders to form strata, we will argue that there is a better way to use the partial information from the sibs. For ease of exposition, we will assume that the elders take on distinct values, although this is not necessary for the following. Let $y_{(1)},\ldots,y_{(n)}$ be the order statistic of the elders. Consider an elder, $y_{(i)}$, where $1 < i < n$ and one of its sibs, which was ranked larger by the expert. Under the Polya posterior, i.e., the assumption that, given the sample, the only y values that can occur in the population are those that have appeared in the sample, this sib must assume one of the $n - i$ values in the set $\{y_{(i+1)},\ldots,y_{(n)}\}$. Similarly, if the expert ranked the sib as smaller than the elder, the sib must take one of the $i - 1$ values in the set $\{y_{(1)},\ldots,y_{(i-1)}\}$. In the case that $y_{(n)}$ has a sib ranked larger than it, we will adopt the convention that it will be assigned the value $y_{(n)}$. Similarly, a sib ranked smaller than $y_{(1)}$ will be assigned this value.

Therefore, given a judgment sample, we see that the Polya posterior can use this partial information about the sibs in a natural way. Given the y values of the elders and the order information about all of the sibs, one first uses Polya sampling to simulate possible values for all of the sibs, but in such a way that the order restrictions for all sibs are satisfied. Once this is done, one can use the observed values of the elders, the simulated values for the sibs, and Polya sampling to simulate values for the rest of the units in the population. This is a natural extension of the Polya posterior to the judgment sampling setup. We will call this restricted version of the Polya posterior the constrained Polya posterior (CPP).

3.3 Some Simulations

To examine how our approach performs, we will now present some simulation results. We will begin by simulating from a population that we constructed. In this population, the y_i values are a random sample of size 2,000 from a gamma population with shape parameter 7 and scale parameter 1, in which each value is increased by 50. The mean of this population is 57.08.

If we assume that the expert is never wrong with their ordering, then when performing simulations, we can use the y values of the sibs to produce the expert's orderings when sibs are compared to their elders. However, this is unlikely to always be true, so it is of interest to observe how a method performs when the expert is fallible. Rather than trying to model how an expert can err, we generated values for an auxiliary variable that is correlated with y. We assume that an expert's orderings are based on this auxiliary variable, for example, x, which is known for all units in the sample. That is, the expert decides that a sib is larger than its elder when the sib's x value is larger than the x value of its elder. By varying the correlation between x and y, one can model different levels of expertise. For our example, we begin by considering two different choices for the x variable. For the first, $x1$, for unit i, we let $x1_i$ given y_i be normally distributed with mean y_i and standard deviation 1.5; all of these distributions are independent. The correlation between $x1$ and y is 0.88. The variable $x2$ is generated in the same way, except the standard deviation is 4. The correlation between $x2$ and y is 0.57. For our purposes, $x1$ represents an expert who will be correct most of the time, whereas $x2$ represents an expert who is not much better than a coin flip.

For the first set of simulations, we let n, the number of elders in the sample, be 10 and consider three subcases in which $m-1$, the number of sibs for each elder, is 2, 5, and 10. In addition, for each case, we consider three subcases in which the expert's ranks are based on y, $x1$, and $x2$, respectively. In each case, we observe 500 samples; for each sample, we simulate 1,000 complete copies of the population to compute the estimator based on CPP.

We compared this estimator to the JPS estimator and the estimator proposed by Frey and Feeman (2012), which we denote with "F-F." We also computed the sample mean of the elders. All of the estimators were approximately unbiased. The results are given in table 1, which shows the average absolute errors for all, except for the JPS estimator. We did not include these results because the F-F estimator always performed slightly better, which was as expected, given the theoretical results of Frey and Feeman (2012).

The CPP is clearly the best estimator, except for the case in which the expert's rankings are based on $x2$; then, both the CPP and F-F estimators essentially perform like the sample means of the elders. This is not surprising because, in this case, the expert's rankings are primarily noise.

Perhaps it is surprising that the CPP estimator performs significantly better than the F-F estimator. It also performs better as the number of sibs increases,

Table 1. Average Absolute Error When the Rankings for the Judgment Sample Estimators are Based on y, $x1$, and $x2$

	$m = 2$	$m = 5$	$m = 10$
Ranks based on y			
Elders	0.72	0.68	0.67
CPP	0.52	0.44	0.38
F-F	0.57	0.58	0.61
Ranks based on $x1$			
Elders	0.66	0.65	0.64
CPP	0.57	0.51	0.50
F-F	0.62	0.61	0.64
Ranks based on $x2$			
Elders	0.67	0.66	0.67
CPP	0.66	0.65	0.64
F-F	0.69	0.68	0.69

NOTE. The results are based on 500 samples.

which is not the case for the F-F estimator. Both, however, perform less well when the expert's ranks are based on $x1$ rather than y.

We created a fourth x variable by using the same set of y values, where for unit i we let $x4_i$, given y_i, be normally distributed with mean $0.8y_i$ and standard deviation 1.5. In this case, our expert was biased downward. The correlation between y and $x4$ was equal to 0.83. In our simulation, we set $n = 10$ and considered the three cases in which $m - 1$ was 2, 5, and 10. In these cases, the average absolute errors for estimating the population mean by using the CPP estimator were 0.61, 0.53, and 0.52; again, these were very similar to the results based on $x1$. Therefore, it is clear that a regular downward bias in the expert does not affect our estimator. The same is true for the JPS and F-F estimators.

We generated another population of values for the x variable. As before, the conditional distribution of the x value, given y_i, was normal with mean y_i, but the conditional standard deviation was 10. This resulted in a correlation of 0.24 between x and y. We set $n = 10$ considering the cases in which $m - 1 = 2$, 5, and 10. As before, we generated 500 samples for each case. For these simulations, the CPP and F-F estimators perform almost the same and just slightly poorer than inferences based on just the elders. This remains true even when the x variable is uncorrelated with y.

We considered other choices for n, the number of elders in the sample, and other choices for generating the number of sibs and other populations of x and y. However, we focused on examples with a small sample size for the elders because these are the kind of examples that typically arise in JPS

sampling. In all cases, the results were very similar to those presented here. When the rankings of the expert are based on real knowledge, the CPP estimator always produces better results than the JPS estimators. The better the expert, the better is the performance of the CPP estimator.

3.4 Why is the CPP Estimator Better?

For each simulated complete copy of the population and for each elder $i \in s$, let p_i denote the proportion of units that assume the value y_i. Then, $\sum_{i\in s} p_i = 1$ and the simulated population mean is $\sum_{i\in s} p_i y_i$. Because our estimate of the population mean is the average of many such simulated population means, it can be written as $\sum_{i\in s} E(p_i)y_i$, where the expectation is made with respect to the CPP. If we let $w_i = nE(p_i)$, our estimator is given by

$$\delta_w = \sum_{i\in s} \frac{w_i}{n} y_i \tag{3}$$

where the w_i values can only be found through simulation and they must sum to n, the number of elders in the sample. The sample mean of the elders is an estimate of this form where the $w_i's \equiv 1$. Under the CPP, the w_i values can be quite different from 1, depending on the ranking information contained in the sibs. The JPS estimator is also of this form and makes use of the ranking information from the sibs to reweight the elders instead of giving equal weight to all. Comparing the weights of the JPS estimator to those of the CPP estimator will help us to determine how the CPP estimator makes more efficient use of the ranking information than the JPS estimator.

Consider a case in which we have five elders and each elder has four sibs associated with it. Suppose in the sample that each elder receives the same ranking. For example, each might be the second smallest in its group, or perhaps each is larger than all of its sibs. Now it is easy to verify that, in each case, the F-F estimate is just the sample mean of the elders. Clearly, this is not sensible. Why estimate the sample mean of the elders when each elder was the largest in its group? How does the CPP estimate make use of this information?

Let $y_{(1)},\ldots,y_{(5)}$ be the order statistic of the five elders in the sample. For a simulated complete copy of the population, let $p_{(i)}$ be the proportion of units that take the value $y_{(i)}$. The CPP estimate is $(1/5)\sum_{i=1}^{5} w_{(i)}y_{(i)}$, where $w_{(i)}$ is five times the expected value of $p_{(i)}$ under the constrained Polya posterior. Table 2 provides the approximate values (found by simulation) of the $w_{(i)}$ values for the five cases in which all the elders have the same rank.

The CPP estimate does something much more sensible than the JPS estimate. For example, when each of the elders is the smallest member of their group, the estimate gives most of the weight to the largest of the elders and much less weight to the rest. In the case of five elders, each with four sibs and

Table 2. Weights Assigned to the Elders in the CPP Estimate When the Ranks of the Elders are All the Same for a Sample of Five Elders, Each with Four Sibs

Ranks	$w_{(1)}$	$w_{(2)}$	$w_{(3)}$	$w_{(4)}$	$w_{(5)}$
All 1	0.20	0.23	0.36	0.54	3.67
All 2	0.89	0.49	0.40	0.52	2.70
All 3	1.84	0.48	0.43	0.48	1.77
All 4	2.71	0.53	0.39	0.38	0.99
All 5	3.82	0.34	0.27	0.37	0.20

NOTE. In each case, the results are based on 5,000 simulated full copies of the population.

Table 3. Average Weights Assigned to the Elders in the CPP and F-F Estimates for 500 Samples with Five Elders, Each with Four Sibs

Method	$y_{(1)}$	$y_{(2)}$	$y_{(3)}$	$y_{(4)}$	$y_{(5)}$
Ranks based on y					
CPP	1.33	0.80	0.75	0.80	1.32
F-F	1.06	0.97	0.96	0.95	1.06
Ranks based on $x1$					
CPP	1.34	0.80	0.77	0.79	1.31
F-F	1.06	0.97	0.96	0.96	1.05

in which each rank appears exactly once, the F-F estimate is the sample mean of the elders. For the case in which for $i = 1,. . .,5$, the rank of $y_{(i)}$ is i, this makes sense, and a simulation suggests that the CPP estimate agrees with the F-F estimate and is approximately the sample mean. Instead, suppose that the rank of $y_{(i)}$ is $6 - i$; in this case, it is not as clear how to weight the five elders. In this case, a simulation found that, approximately, the values for the five weights in the CPP estimate were 2.11, 0.27, 0.24, 0.27, and 2.11.

To further explore this question, using the constructed population in Section 3.3, we examined 500 random samples of five elders in which each had four sibs and considered two cases. In the first case, we used the y values to construct the rankings; in the second, we used the $x1$ values. In each case, we found the average weights that the CPP estimates and the F-F estimates assigned to the five elders in the sample. The results are given in table 3. On average, the F-F weights are closer to uniform, whereas the CPP places more weight on the extremes and less in the middle. This pattern generally holds, but as the number of elders increases and the number of sibs increases, the two averages tend to be more similar, although it continues to be true that the CPP assigns more weight to the extremes.

4. THE STEPWISE BAYES MODEL AND COMPUTING OUR ESTIMATOR

In the first part of this section, we will briefly outline the theory underlying our estimator. One can show that this estimator is admissible for estimating the population mean under squared error loss. This argument is a generalization of the one given in Ghosh and Meeden (1997), which proves the admissibility of the sample mean and is based on the stepwise nature of these estimators. We start by presenting a true Bayesian model, which leads to the sequence of stepwise Bayes models that yield our estimator.

For ease of exposition, we assume that the units in the population can take only finitely many values, $b = (b_1,\ldots,b_r)$, for some positive integer r. We assume that these values are known a priori and labeled in such a way that $b_i < b_{i+1}$ for $i = 1,\ldots,r-1$. We begin by assuming a Bayes model for the population y, which is a Dirichlet mixture of independent multinomial random variables. Let $\theta = (\theta_1,\ldots,\theta_r)$ belong to the $r-1$ dimensional simplex, Θ. That is, each $\theta_i \geq 0$ and $\sum_{i=1}^{r} \theta_i = 1$. Then, for some $\varepsilon > 0$, the probability model for the population is

$$\begin{aligned} &\theta \sim \text{Dirichlet}(\varepsilon,\ldots,\varepsilon) \\ &y_i's|\theta \sim \text{iid multinomial}(1,\theta) \end{aligned} \tag{4}$$

For a population vector y, let $c_i(y)$ be the number of times that b_i occurs in y; also, $c_i(y) \geq 0$ and $\sum_{i=1}^{r} c_i(y) = N$. Under our model, we have

$$\begin{aligned} p(y) &= \frac{\Gamma(r\varepsilon)}{\Gamma(\varepsilon)^r} \int \cdots \int_{\Theta} \prod_{i=1}^{r} \theta_i^{c_i(y)+\varepsilon-1} d\theta \\ &= \frac{\Gamma(r\varepsilon)}{\Gamma(\varepsilon)^r} \frac{\prod_{i=1}^{r} \Gamma(c_i(y)+\varepsilon)}{\Gamma(N+r\varepsilon)} \end{aligned}$$

Let *set* denote some subset of $1,\ldots,N$. Then, it follows that

$$p(y(set)) = \frac{\Gamma(r\varepsilon)}{\Gamma(\varepsilon)^r} \frac{\prod_{i=1}^{r} \Gamma(c_i(y(set))+\varepsilon)}{\Gamma(N_{set}+r\varepsilon)}$$

where $c_i(y(set))$ denotes the number of times the value b_i appears in $y(set)$ and N_{set} is the number of units in the population that belong to *set*. In the following, *eld* will denote the set of labels for the elders in the sample, *sib* will denote the labels for the set of sibs in the sample, and *uns* will denote the labels for the rest of the population.

Given a sample, let Δ denote the set of all possible $y(sib)$ values that satisfy the constraint information in the sample produced by the expert. Let $y(sib)$

denote a member of Δ. So, given $y(sib) \in \Delta$, we are interested in

$$\begin{aligned} p(y(sib), y(uns)|y(eld), y(sib) \in \Delta) &= \frac{p(y(eld), y(sib), y(uns))}{p(y(eld)), ysib \in \Delta)} \\ &= p(y(sib)|y(eld), y(sib) \in \Delta) p(y(uns)|y(eld, sib)) \\ &= \frac{p(y(eld, sib))}{\sum_{y(sb)' \in \Delta} p(y(eld), y(sib)')} p(y(uns)|y(eld, sib)) \\ &= \frac{\prod_{i=1}^{r} \Gamma(c_i(y(eld, sib)) + \varepsilon)}{\sum_{y(sb)' \in \Delta} \left(\prod_{i=1}^{r} \Gamma(c_i(y(eld)) + c_i(y(sib)' + \varepsilon)) \right)} \times p(y(uns)|y(eld, sib)) \end{aligned}$$

The stepwise Bayesian approach to proving admissibility allows one to assume that the unique values in $y(eld)$ are exactly the values of b. This justifies setting $\varepsilon = 0$ in the last line of the previous equation to obtain

$$\begin{aligned} &p(y(sib), y(uns)|y(eld), y(sib) \in \Delta) \\ &= p(y(sib)|y(eld), y(sib) \in \Delta) p(y(uns)|y(eld, sib)) \end{aligned} \tag{5}$$

where

$$\begin{aligned} &p(y(sib)|y(eld), y(sib) \in \Delta) \\ &\qquad = \frac{\prod_{i=1}^{r} \Gamma(c_i(y(eld, sib)))}{\sum_{y(sb)' \in \Delta} \left(\prod_{i=1}^{r} \Gamma(c_i(y(eld)) + c_i(y(sib)')) \right)} \end{aligned} \tag{6}$$

As before, we adopt the convention that when the expert asserts that a sib value is less than b_1, then that sib is assigned the value b_1. Similarly, if the expert asserts that a sib value is greater than b_r, then that sib is assigned the value b_r.

The first term on the right-hand side of (5) is just the constrained version of the Polya posterior for $y(sib)$ given $y(eld)$ and the information contained in Δ; the second term is just the Polya posterior of $y(uns)$ given $y(eld)$ and $y(sib)$. Therefore, this is the posterior distribution that was outlined in Section 3.2. It is not, however, a true posterior distribution arising from a single prior distribution that was specified before the sample was observed.

When n/N is small, rather than using Polya sampling from an urn to simulate full copies of a population, it is more efficient to use the Dirichlet approximation to the Polya distribution. Consider an urn that contains $v_i \geq 1$ balls of type i for $i = 1,\ldots,n$. Let $p = (p_1,\ldots,p_n)$, where p_i is the proportion of balls of type i in the urn after a large number of Polya draws. The distribution of p is approximately Dirichlet $(v_1,\ldots,v_n)$. In our problem, this means that when n/N

is small and we are estimating the population mean, we do not need to know the population size.

For small problems, the sum in the denominator in (6) is easy to compute, but for larger problems, this becomes problematic. Therefore, although the basic idea underlying our approach is simple, the computation of our estimate is more complicated.

One might try rejection sampling to simulate from $p(y(sib)|y(eld),y(sib) \in \Delta)$, but in practice this will not work well. Instead, we will use importance sampling to compute our estimates. Given a sample, we will first randomly assign values to the sibs that are consistent with the information about their order. Once all the sibs have been assigned a value, we use Polya sampling to simulate values for all of the unsampled units in the population. We will implement this process many times to generate a large number of simulated copies of the population. For each such copy of the population, we will find its mean. Our estimate will just be the appropriate weighted average of these means; the weights are determined by our importance sampling distribution.

We now describe this process in more detail. Suppose we have a judgment sample; that is, the y values of the elders and the order information of their associated sibs. As we have explained, for any sib, there is a subset of possible y values that can be assigned to this sib, which is consistent with the observed y values of the elders and the order information for the sib. For each sib, we randomly select one of these values and assign it to the sib. This is done independently across the sibs and the joint probability of these assignments is the same for all possible assignments. We denote this value by λ. The numerator of (6), up to the normalizing constant, is the joint probability of these assignments to the sibs under the constrained Polya posterior. The ratio of this latter value to λ will be proportional to the weight attached to the simulated complete copy of the population, based on this assignment of values to the sibs. If $y(sib)_j$ denotes the jth simulated set of values for the sibs under our importance sampling distribution, its importance weight, up to a constant, is given by

$$iw_j = \prod_{i=1}^{r} \Gamma(c_i(y(eld)) + c_i(y(sib)_j))$$

Let $\tilde{\mu}_j$ denote the mean of the simulated complete copy of the population made up of the values in $y(eld)$, $y(sib)_j$, and $y(uns)_j$. Assume that we have generated K such simulated complete copies of the population; then, our estimate of the population mean will be

$$\sum_{j=1}^{K} iw_j \tilde{\mu}_j / \sum_{j=1}^{K} iw_j$$

A referee suggested that a possible problem with the Polya posterior is that it assumes that the only values that can appear in the population are those that

have appeared in the sample. We would like to point out that many of the standard estimators in survey sampling implicitly assume that any unobserved unit in the population must take one of the values observed in the sample. This is particularly true for the sample mean and more generally true for estimators like the Horvitz-Thompson, which attach weights to units in the sample to compute an estimate. Ghosh and Meeden (1997) give an argument for proving the admissibility of the sample mean for estimating the population mean that is based on the stepwise Bayes nature of the sample mean. This argument uses the Polya posterior, which explicitly builds on the assumption that the unseen units can only have values that have appeared in the sample. This technical argument dovetails nicely with the underlying intuition behind these estimators, although explicitly saying this can seem to be surprising. A mild generalization of that argument can be used to prove the admissibility of the estimator under consideration here which takes into account the partial information about the sibs from an expert who is always correct. This argument also works for experts who announce ties and who use greater than or equal to, or less than or equal to, as a possible ordering.

5. FINDING INTERVAL ESTIMATORS

Given several rank ordered samples, one can find a confidence interval for a mean, but for a single rank ordered sample or a single JPS sample, there are no known methods for producing a confidence interval. The standard Bayesian approach for constructing confidence intervals for the population mean by using the CPP would be to generate many, K, simulated complete copies of the population and find the simulated population mean in each case. One finds the lower 0.025 quantile and the upper 0.975 quantile of these K simulated means and uses them to form a 0.95 Bayesian credible interval.

Simulations have shown that for the problems considered here, the CPP will produce intervals that are too short. That is, their frequency of containing the true population mean can be much less than 0.95. This problem with the CPP was noted by Strief and Meeden (2013). This difficulty arises because with a small sample size, as often happens in a judgment sample, and with several constraints, there is just not enough variability in the CPP to produce approximate 95% confidence intervals.

To overcome this problem, they introduced the weighted Dirichlet posterior (WDP). Let n denote the number of elders in the sample; for $i = 1,\ldots,n$, let $w_i = nE(p_i)$, where the expectation is taken under the CPP and p_i is the proportion of a simulated complete copy of the population that assumes the value y_i. The WDP is a Dirichlet distribution with parameter vector $w = (w_1,\ldots,w_n)$ and one can generate possible simulated populations from this distribution, defined by $p = (p_1,\ldots,p_n)$'s drawn from this Dirichlet distribution.

Strief and Meeden (2013) recommended a two-step procedure for finding an interval estimate of a population parameter. First, use the CPP to find the vector of weights $w = (w_1, \ldots, w_n)$ for the elders in the sample, in which $\sum_{i=1}^{n} w_i = n$, the sample size. Next, use the WDP to generate many simulated copies of the population and proceed in the usual Bayesian way. This is easy to do and will result in more variability in the population parameter of interest. If the parameter of interest is the population mean, $\sum_{i=1}^{n} p_i y_i$, one can write down its variance explicitly and not actually conduct the second simulation. One can compute the standard normal theory interval. This is what we will do for our simulations. Because we can simulate complete copies of the population, we can estimate population quantities other than the population mean; in the following, we will also estimate the population median. When estimating the population median, however, one needs to conduct the second simulation.

Further reflection, however, indicates that we must modify the Strief and Meeden (2013) approach for our problem. The CPP point estimator becomes more precise as the number of sibs increases. This means that taking the sum of the elements of the vector w to be n, the number of elders, will result in intervals that are too long. Instead, we must increase this sum by multiplying w by an appropriate constant larger than 1. For a given problem, the appropriate constant will depend on three things: the number of elders, the number of sibs, and the talent of the expert. In practice, the last item can be difficult to gauge correctly.

This problem is even more difficult because the amount of information contained in a sib can vary. To show this, let $y_{(1)}, \ldots, y_{(n)}$ be the order statistic of the n elders in the sample. Now consider a sib of $y_{(1)}$. Learning that this sib is less than $y_{(1)}$ is much more informative than learning that it is greater than $y_{(1)}$. We will give the sib in the first case a score of 1, whereas in the second case, we will give it a score of $1/(n-1)$ because it can be any one of the $n-1$ larger values. More generally, for $1 < i < n$, a sib of $y_{(i)}$ that is less than this will be given a score of $1/(i-1)$, whereas one that is greater will be assigned a score of $1/(n-i)$. Finally, a sib less than $y_{(n)}$ is assigned the score $1/(n-1)$, whereas one greater than this is assigned the score 1. The score of a sib can only be determined once the sample has been observed and can range from $1/(n-1)$ to 1. We denote a sib's score by v_{sib} and note that it is a crude measure of how much information the sib contains. If we let $sibwt = \sum_{sibs} v_{sib}$, then we have a measure of how much additional information is contained in the rankings of all of the sibs.

With this definition, and for $n > 4$, we now define the effective sample size as

$$effsmpsz = n + \left(\frac{1}{2} - \frac{2}{n}\right) sibwt \tag{7}$$

This is used in the second step of the inferential process when finding an approximate 95% confidence interval. After the CPP has been used to find the vector of weights $w = (w_1, \ldots, w_n)$, which sums to n, the number of elders in the sample, the elements of w are each multiplied by $effsmpsz/n$ so that the re-scaled vector now sums to $effsmpsz$. This rescaled vector is used in the WDP when finding the interval estimate. We believe that the preceding definition makes intuitive sense and it works reasonably well in some simulation studies. We believe that it yields a sensible approximate solution for many situations in which an excellent expert is available. Here, we have been assuming that the expert is very good and makes few mistakes. For poorer experts, one needs to decrease the second term in the right-hand side of (7). How much it must be decreased depends on the expert.

Before presenting some simulation results, we note that there is nothing in our approach that precludes different elders from having different numbers of sibs associated with them. To see what may happen, we returned to the constructed population in Section 3.3. Now, instead of assigning each elder the same number of sibs, we allowed the number of sibs be a Poisson random variable. We considered three cases in which the means of the Poisson random variables were 2, 5, and 10, and we based the expert's rankings on $x1$. The results for both the mean and median of the population are given in table 4.

Table 4. Comparison of Inferential Methods When the Number of Elders is 10, the Number of Sibs is Random, and the Expert Rankings are Based on $x1$

Parameter	Method	Average value	Average absolute error	Average length	Frequency of coverage
	Average number of sibs is 2				
Mean	Elders	57.13	0.72	3.15	0.884
	WDP	57.11	0.62	3.09	0.928
Median	Elders	56.91	0.80	4.37	0.928
	WDP	56.55	0.65	3.96	0.944
	Average number of sibs is 5				
Mean	Elders	57.07	0.67	3.27	0.918
	WDP	57.11	0.53	3.02	0.964
Median	Elders	56.82	0.78	4.48	0.954
	WDP	56.50	0.59	3.81	0.974
	Average number of sibs is 10				
Mean	Elders	57.08	0.62	3.18	0.932
	WDP	57.10	0.45	2.67	0.952
Median	Elders	56.87	0.72	4.30	0.956
	WDP	56.51	0.54	3.33	0.960

NOTE. The true population mean and median are 57.08 and 56.82, respectively. The results are based on 500 samples.

The average absolute errors for estimating the population mean were very similar to those in table 1. Although we did not include the performance of the interval estimators when the number of sibs was fixed in table 1, those results are similar to those in table 4 for both the population mean and median. Also, there is some evidence that our intervals may be slightly too long for the population median. We ran this simulation again when the expert's rankings were based on y. The behavior of our point estimator did not change much, so table 5 simply provides the length and frequency of coverage for our approximate 0.95 credible interval. Again, there is some evidence that our intervals are too long and contain the true population parameters more than 95% of the time. Another simulation was conducted in which the number of elders was 7 and the average number of sibs was either 3 or 6. The results are summarized in table 6, which shows that the coverage probability seems to be just about right. We also ran other simulations in which the number of sibs was fixed and had the same value for all elders. The behavior of our intervals in these simulations was very similar to those discussed previously in which the number of sibs was random. Overall, our intervals are behaving reasonably, although they can be conservative in some cases.

6. SIMULATION USING THE SELLING PRICES OF HOUSES

Our population is the set of houses sold between November 1, 2011, and October 31, 2012, in two zip code areas in Saint Paul, Minnesota. There are 597 such houses. The y value is the sale price of the house. In addition, the x variable is the real estate tax for the house at the time of sale. The correlation between these two variables is 0.86. Given a sample, we are interested in estimating either the average sale price or the median sale price of this population

Table 5. Performance of the WDP Intervals When the Number of Elders is 10, the Number of Sibs is Random, and the Expert Rankings are Based on y

Mean		Median	
Average length	Frequency of coverage	Average length	Frequency of coverage
Average number of sibs is 2			
3.11	0.960	3.77	0.958
Average number of sibs is 5			
2.94	0.986	3.58	0.972
Average number of sibs is 10			
2.62	0.966	3.18	0.972

NOTE. The results are based on 500 samples.

Table 6. Performance of the WDP Intervals When the Number of Elders is 7 and λ is the Average Number of Sibs

λ	Mean		Median	
	Average length	Frequency of coverage	Average length	Frequency of coverage
	Rankings based on y			
3	3.65	0.962	4.24	0.954
6	3.45	0.942	3.99	0.942
	Rankings based on $x1$			
3	3.63	0.954	4.19	0.964
6	3.47	0.942	4.16	0.930

NOTE. The results are based on 500 samples.

of homes. We consider two cases in which the expert's rankings can be based either on x, the real estate tax for the house at the time of sale, or y, the true sale price. We consider three different cases in which the number of elders was 10, but the number of sibs for each elder was a Poisson random variable with means, 2, 5, and 10, respectively. For each case, we generated 500 samples and compared the WDP point and interval estimates for the population mean and median to the estimates based on the elders. The results are given in table 7. The WDP point estimates behave in a sensible fashion and are always better than those based on just the elders. The ones based on y perform better than those based on x and tend to become better as the number of sibs increases. The point estimates for the median appear to be a bit biased downward. The interval estimates for the mean are close to the nominal coverage of 0.95, whereas the intervals for the median seem to overcover slightly.

We repeat the simulation when the number of elders is 7 and the number of sibs is a Poisson random variable with a mean of either 3 or 6. For estimating the population mean, the frequency of coverage of our intervals are 0.902 and 0.914 when the ranks are based on x and 0.926 and 0.932 when the ranks are based on y. For estimating the population median, these numbers are 0.946 and 0.960, and 0.958 and 0.954, respectively. Overall the simulations indicate that our stepwise Bayesian credible intervals will cover the true parameter value approximately 95% of the time for the small sample sizes that often occur in rank set sampling. The actual coverage probability will depend, in part, on the accuracy of the expert. This may be difficult to determine in practice.

7. FINAL REMARKS

The sample mean is a sensible estimator of the population mean if, given a sample, one believes that the sampled and unsampled units are roughly

Table 7. Comparison of Inferential Methods for the Real Estate Population When the Number of Elders is 10, the Number of Sibs is Random, and the Expert Rankings are Either Based on *x*, the Amount of Taxes Paid, or *y*, the True Sale Price

Parameter	Average number of sibs	Method	Average value	Average absolute error	Average length	Frequency of coverage
Mean		Elders	265.0	32.80	157.3	0.896
Median		Elders	239.6	29.53	184.0	0.962
Mean	2	WDP-*x*	267.1	29.34	154.8	0.936
		WDP-*y*	265.6	26.85	151.6	0.940
Median		WDP-*x*	228.6	23.87	164.8	0.966
		WDP-*y*	228.3	21.03	155.4	0.972
Mean	5	WDP-*x*	267.7	27.70	146.2	0.928
		WDP-*y*	264.1	22.40	142.3	0.946
Median		WDP-*x*	226.8	22.78	155.6	0.964
		WDP-*y*	225.5	17.81	144.3	0.970
Mean	10	WDP-*x*	268.9	24.64	129.9	0.948
		WDP-*y*	265.9	19.51	127.9	0.958
Median		WDP-*x*	229.5	19.23	133.2	0.962
		WDP-*y*	227.1	14.12	125.5	0.974

NOTE. The mean and median of the population sale prices, in thousands of dollars, are 266.2 and 235. The results are based on 500 samples.

exchangeable. In the design-based approach, this belief follows from the fact that the sampling design was simple random sampling without replacement. In this framework, the JPS estimators incorporate the additional knowledge contained in the ranks of the sibs by stratifying the elders based on their ranks within their cohorts of sibs. The Polya posterior is another way to model exchangeability between the observed and unobserved. Although not a pure Bayesian posterior, it has a stepwise Bayes justification that yields a noninformative Bayes justification of the sample mean as a good estimator of the population mean. It can include information about the sibs in the usual Bayesian fashion by considering them as observations that have been censored at random in a particular way. The CPP point estimates for the population mean are clearly superior to the JPS estimators because they make more effective use of the information contained in the ranking of the sibs.

A referee asked if it would be possible to extend our approach to situations in which more than one expert is performing the ranking. We believe that the answer is yes. For the case in which the two experts have similar abilities, there should be no problem. There is nothing in our approach that assumes that only one expert is doing the ranking. In fact, this is most likely the case in our

real estate example. However, this suggests the following problem. Suppose that we have two experts: the first is better than the second, but is also more costly. How should we allocate the number of items we ask each of them to rank?

The CPP is more flexible than the JPS because the number of sibs associated with each elder can vary. In fact, there may be some elders with no sibs at all. In addition, the expert needs only to compare each sib to its elder and not rank sets of units. This yields sensible interval estimates for the population mean and point and interval estimates for the population median, although more study needs to be undertaken to calibrate the interval estimators for the median. This method can account for ties and rankings that allow for possible equality for the two units under comparison. As a final application, suppose that the values of the y variable are qualitative, but with a natural order. One can use a sample of elders with orderings of the sibs to estimate the proportion of units in the population of each type.

Supplementary Materials

The Sweave document (Meeden [2013]) contains *R* code that, given a judgment sample, approximately finds the w_i/n values of (3). It also finds the *effsmpsz* given in (7) and the corresponding WDP variance of the population mean. This should make it easy for anyone familiar with the computer package *R* (R Core Team [2013]) to compute our estimator.

REFERENCES

Binder, D. (1982), "Non-parametric Bayesian Models for Samples from a Finite Population," *Journal of the Royal Statistical Society, Series B*, 44, 388–393.

Booth, J. G., R. W. Butler, and P. Hall (1994), "Bootstrap Methods for Finite Population Sampling," *Journal of the American Statistical Association*, 89, 1282–1289.

Chen, Z., Z. Bai, and B. K. Sinha (2004), *Ranked Set Sampling: Theory and Applications*, New York: Springer-Verlag.

Frey, J., and T. G. Feeman (2012), "An Improved Mean Estimator for Judgment Poststratification," *Computational Statistics and Data Analysis*, 56, 418–426.

Ghosh, M., and G. Meeden (1997), *Bayesian Methods for Finite Population Sampling*, London: Chapman and Hall.

Gross, S. (1980), "Median Estimation in Survey Sampling," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 181–184.

Hartley, H. O., and J. N. K. Rao (1968), "A New Estimation Theory for Sample Surveys," *Biometrika*, 55, 159–167.

Lazar, R., G. Meeden, and D. Nelson (2008), "A Noninformative Bayesian Approach to Finite Population Sampling Using Auxiliary Variables," *Survey Methodology*, 34, 51–64.

Lo, A. (1988), "A Bayesian Bootstrap for a Finite Population," *Annals of Statistics*, 16, 1684–1695.

MacEachern, S. N., E. A. Stasny, and D. A. Wolfe (2004), "Judgment Poststratification with Imprecise Rankings," *Biometrics*, 60, 207–215.

McIntyre, G. A. (1952), "A Method for Unbiased Selective Sampling Using Ranked Sets," *Australian Journal of Agricultural Research*, 3, 385–390.

Meeden, G. (2013), "Computing Constrained Polya Posterior Estimates when Using Judgment Sampling," University of Minnesota School of Statistics, Technical Report. Accessed November 1, 2013, from http://purl.umn.edu/157317.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna. Accessed November 1, 2013, from http://R-project.org.

Rubin, D. (1981), "The Bayesian Bootstrap," *Annals of Statistics*, 9, 130–134.

Strief, J., and G. Meeden (2013), "Objective Stepwise Bayes Weights in Survey Sampling," *Survey Methodology*, 39, 1–27.