Survey Methodology, June 1995
Vol. 21, No. 1, pp. 71–77
Statistics Canada

Median Estimation Using Auxiliary Information

GLEN MEEDEN[1]

ABSTRACT

The problem of estimating the median of a finite population when an auxiliary variable is present is considered. Point and interval estimators based on a non-informative Bayesian approach are proposed. The point estimator is compared to other possible estimators and is seen to perform well in a variety of situations.

KEY WORDS: Sample survey; Estimation; Median; Auxiliary variable; Quantile; Non-informative Bayes.

1. INTRODUCTION

The problem of estimating a population mean in the presence of an auxiliary variable has been widely discussed in the finite population sampling literature. The ratio estimator has often been used in such situations. For the problem of estimating a population median the situation is quite different. Only recently has this problem been discussed. Chambers and Dunstan (1986) proposed a method for estimating the population distribution function and the associated quantiles. They assumed that the value of the auxiliary variable was known for every unit in the population and their estimator came from a model-based approach. Rao *et al.* (1990) proposed ratio and difference estimators for the median using a design-based approach. Kuk and Mak (1989) proposed two other estimators for the population median. To use the Kuk and Mak estimators one only needs to know the values of the auxiliary variable for the units in the sample and its median for the whole population. The efficiencies of these estimators depend directly on the probability of 'concordance' rather than on the validity of an assumption of linearity between the variable of interest and the auxiliary variable.

Recently Meeden and Vardeman (1991) discussed a non-informative Bayesian approach to finite population sampling. This new approach uses the 'Polya posterior' as a predictive distribution for the unobserved members of the population once the sample has been observed. Often it yields point and interval estimates that are very similar to those of standard frequentist theory. Moreover it can be easy to implement in problems that are difficult for standard theory. In this note we show how this method can be used for the problem of estimating a population median when an auxiliary variable is present and compare it to some of the other proposed methods.

2. ESTIMATING THE MEDIAN

Consider a finite population containing N units. For the unit with label i let y_i denote the characteristic of interest and x_i the auxiliary variable. We assume that both y_i and x_i are real numbers and each is known for every unit in the population. Let s denote a typical sample of size n which was chosen by simple random sampling without replacement. We assume simple random sample for convenience, since in many problems of this type the sampling will often be more purposeful. Before considering the problem of estimating the median of the population we review some well known facts about the problem of estimating the mean.

Consider the super population model where it is assumed that for each i, $y_i = bx_i + u_i e_i$. Here b is an unknown parameter while the u_i's are known constants and the e_i's are independent identically distributed random variables with zero expectations. Since the population mean can be written as $N^{-1}(\sum_{i\in s} y_i + \sum_{j\notin s} y_j)$ we would expect $N^{-1}(\sum_{i\in s} y_i + \hat{b}\sum_{j\notin s} x_j)$ to be a sensible estimate of the mean whenever $\hat{b}$ is a sensible estimate of b. One particular choice of $\hat{b}$ is the weighted least squares estimator where the weights are determined by the u_i's. For example if for all i, $u_i = \sqrt{x_i}$, the resulting estimator is just the usual ratio estimator. While if for all i, $u_i = x_i$, then $\hat{b} = n^{-1}\sum_{i\in s}(y_i/x_i)$ and the resulting estimator is one that was discussed by Basu (1971). (See also Royall (1970).) Using this super population setup it is easy to generate populations where the ratio estimator has smaller mean squared error than the Basu estimator and vice versa. A somewhat limited simulation study on a variety of populations found that the performance of the Basu estimator is quite similar to the performance of the ratio estimator although in the majority of the cases the ratio estimator performs better than the Basu estimator. This is not unexpected, given the wide use of the ratio estimator.

[1] Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

In Meeden and Vardeman (1991) a non-informative Bayesian approach to finite population sampling, based on the Polya posterior, was developed. For the simple problem where no auxiliary variable is present, given the observed values in the sample, it introduces a Polya urn distribution as a pseudo posterior distribution over the unobserved members of the population. This pseudo posterior distribution can be used to obtain point and interval estimates of a variety of population quantities. It is related to the Bayesian bootstrap of Rubin (1981) and the Dirichlet process prior of Ferguson (1973). When estimating the median it yields results similar to those of Binder (1982). A theoretical justification for it is a stepwise Bayes argument which yields the admissibility of the resulting estimators. See for example Meeden and Ghosh (1983). There the admissibility of the Basu estimator was demonstrated. In that case the Basu estimator was shown to arise from a 'posterior' which treats the known and unknown ratios, $r_i = y_i/x_i$ as exchangeable. Note that this is very similar in spirit to the super population model justification for this estimator given above, where the ratios $r_i = y_i/x_i$ were independent and identically distributed. We shall see that the stepwise Bayes logic underlying the Basu estimator for the mean carries over in a straight forward way to point and interval estimators for the median. Unfortunately this is not the case for some of the other estimators. One natural but perhaps naïve estimator which mimics in some sense the ratio estimator of the mean is just the ratio of the median of the y values in the sample to the median of the x values in the sample multiplied by the median of the x values in the population. There is no known model based theory which underlies this estimator as is the case for the ratio estimator of the mean.

In the Bayesian approach to finite population sampling one needs to specify a prior distribution. Then given a sample, inferences are based on the posterior distribution, which is the predictive distribution for the unseen members of the population given the units in the sample. In the stepwise Bayes approach, given the sample one always has a 'posterior' distribution but it does not arise from a single prior distribution. However this 'posterior' distribution can be used in the usual Bayesian manner to find point and interval estimators of parameters of interest. We now will show how the stepwise Bayes model which yields Basu's estimator for the mean can also be used when estimating the median. In this setup, given a sample, the predictive distribution for the unobserved ratios treats the observed and unobserved ratios as 'exchangeable'.

For definiteness suppose our sample contains the first n units of the population. We construct an urn which contains n balls where ball i is given the value of the i-th observed ratio, say r_i. We begin by selecting a ball at random from the urn and the observed value is assigned to the unobserved unit $n + 1$. This ball and an additional ball with the same value is returned to the urn. Another ball is chosen from the urn and its value is assigned to the unobserved unit $n + 2$. This ball and another with the same value are returned to the urn. This process is continued until all of the unobserved units have been assigned a ratio. Once they have all been assigned a value we have observed one realization from our 'posterior' distribution for the unseen ratios given the sample of seen ratios. If in this process the unobserved unit j has been assigned the ratio with value r we then assign its y_j value to be rx_j. Hence using simple Polya sampling we have created a predictive distribution for the unobserved units given the sample. We call this predictive distribution the 'Polya posterior'. It is easy to check that this predictive distribution gives the Basu estimator when estimating the population mean under squared error loss.

Given the sample the 'Polya posterior' yields a predictive distribution for the unobserved members of the population and hence a predictive distribution for the median as well. From the decision theory point of view the usual loss function is absolute error when estimating a median. For this loss function the Bayes estimate is just the median of the posterior or predictive distribution for the population median. If one were using squared error loss for estimating the median then the Bayes estimate is just the mean of the predictive distribution for the population median. The admissibility of these estimators under the appropriate loss function follows from a stepwise Bayes argument in the same way as the proof of admissibility for the Basu estimator of the population mean. In Meeden and Vardeman (1991) and Meeden (1993) the following somewhat surprising fact was noted. For many common distributions the mean of the predictive distribution for the population median performed better than the median of the predictive distribution for the population median under both loss functions. Similar results hold for this problem. Hence our estimator will be the mean of the predictive distribution for the population median even though we will follow standard practice and use absolute error as our loss function. We will denote this estimator by *estpp*. This estimator cannot be found explicitly. However we will find it approximately by simulating observations from the posterior or predictive distribution for the population median. Under the Polya sampling scheme for the ratios described above we can simulate a possible realization of the entire population. For this simulated copy we can then find its median. If we repeat this process R times then we have simulated the predictive distribution of the population median under the 'Polya posterior'. When R is large the mean of these R simulated population medians yields, approximately, the estimate *estpp*.

In what follows we will compare the estimator *estpp* to several other estimators. Another estimator we consider is just the sample median of the y_i's. This ignores the information contained in the auxiliary variable and is used as a bench mark. It will be denoted by *estsm*. Another

estimator is the natural analogue of the ratio estimator of the population mean. This is discussed in Kuk and Mak (1989) and denoted by *estrm*. It is just the ratio of the median of the y values to the median of the x values in the sample multiplied by the median of all the x values in the population. They proposed two other estimators for the median. We will consider just the first one and denote it by *estkm*. This estimator has a plausible intuitive justification and can be found in their paper. Rao, Kovar and Mantel (1990) considered a designed based estimator for the median. We will denote this estimator by *estrkm*. Since this estimator can be time consuming to compute we will find it approximately using a method due to Mak and Kuk (1993). Finally we will consider the estimator proposed in Chambers and Dunstan (1986) and denote it by *estcd*. Actually Chambers and Dunstan propose a whole family of estimators and we will only consider one special case which is appropriate when $u_i = \sqrt{x_i}$ in the super population model described at the beginning of this section. We now briefly outline the argument that leads to their estimator of the median. Let F denote the cumulative distribution function associated with the y values of the population. That is F puts mass $1/N$ on each y_i in the entire population. The first step is to get an estimator of $F(t)$ for an arbitrary real number t. If s denotes our sample of size n then given the sample we can write

$$F(t) = N^{-1}\left\{ \sum_{i \in s} \Delta(t - y_i) + \sum_{j \notin s} \Delta(t - y_j) \right\}$$

where $\Delta(z)$ is the step function which is one when $z \geq 0$ and zero elsewhere. Since the first sum in the above expression is known once we have observed the sample, to get an estimate of $F(t)$ it suffices to find an estimate of the second sum. Now under our assumed super population model the population ratios $(y_i - bx_i)/\sqrt{x_i}$ are independent and identically distributed random variables. Since after the sample s is observed a natural estimate of b is $\hat{b} = \sum_{i \in s} y_i / \sum_{i \in s} x_i$ one could act as if the n known ratios $(y_i - \hat{b}x_i)/\sqrt{x_i}$ for $i \in s$ are actual observations from this unknown distribution. Under this assumption, for a fixed t and a fixed unit j not in the sample s an estimate of $\Delta(t - y_j)$ is just the number of the n known ratios incorporating $\hat{b}$ less than or equal to $(t - \hat{b}x_j)/\sqrt{x_j}$ divided by n. Finally if we sum over all the unobserved units j these estimates of $\Delta(t - y_j)$ we then have an estimate for the second sum in the above expression for $F(t)$ which then yields an estimate of $F(t)$. Once we can estimate $F(t)$ for any t by say $\hat{F}(t)$ then the estimate of the population median is $\inf\{t : \hat{F}(t) \geq 0.5\}$.

3. THE POPULATIONS

We will compare these estimators using several different populations. We begin with three actual populations. The first is a group of 125 American cities. The x variable is their 1960 populations, in millions, while their y variable is the corresponding 1970 populations, again in millions. The second is a group of 304 American counties. The x variable is the number of families in the counties in 1960, while the y variable is the total 1960 population of the county. Both variables are given in thousands. The third population is 331 large corporations. The x variable is their total sales in 1974 and the y variable their total sales in 1975. The sales are given in billions of dollars. We denote these three populations by *ppcities*, *ppcounties* and *ppsales*. For the three populations the correlations are .947, .998 and .997. These populations were discussed in Royall and Cumberland (1981). Our *ppcounties* is similar to their population Counties60 except we have taken the x variable to be the number of families rather than the number of households.

We have also considered six artificial populations. In each case the auxiliary variable x was chosen first and then the y variable was generated from it. In some cases we followed the super population model described at the beginning of the previous section for some choice of the u_i's. In some other cases we violated the assumption that conditional on the value x_i the mean of y_i is bx_i. In all cases the errors, the e_i's, were independent and identically distributed normal random variables with mean zero and variance one.

In the first population, *ppgamma20*, the x_i's were a random sample from a gamma distribution with shape parameter twenty and scale parameter one. Then given x_i the conditional distribution of y_i was normal with mean $1.2x_i$ and variance x_i, *i.e.*, $u_i = \sqrt{x_i}$.

In the second population, *ppgamma5a*, the x_i's were ten plus a random sample from a gamma distribution with shape parameter five and scale parameter one. Then given x_i the conditional distribution of y_i was normal with mean $3x_i$ and variance x_i.

In *ppgamma5b* the auxiliary variable was the same as in *ppgamma5a*. Then given x_i the conditional distribution of y_i was normal with mean $3x_i$ and variance x_i^2.

In *ppstskew* the auxiliary variable was strongly skewed to the right with mean 42.63, median 39.29 and variance 204.59. Then given x_i the conditional distribution of y_i was normal with mean $x_i + 5$ and variance $9x_i$.

In *ppln* the auxiliary variable was a random sample from a log-normal population with mean and standard deviation (of the log) 4.9 and .586 respectively. Then given x_i the conditional distribution of y_i was normal with mean $x_i + 2 \log x_i$ and variance x_i^2.

In *ppexp* the auxiliary variable was fifty plus a random sample from the standard exponential distribution. Then given x_i the conditional distribution of y_i was normal with mean $80 - x_i$ and variance $(.6 \log x_i)^2$.

All the populations contain 500 units except *ppstskew* which has 1,000. The correlations between the two variables for these last six populations are .76, .87, .41, .61, .58 and $-.28$ respectively.

In most examples where ratio type estimators are used both the y_i's and x_i's are usually strictly positive. In population *ppstskew* 13 of the 1,000 units have a y value which is negative. In the original construction of population *ppln* quite a few more of the y values were negative. The population was modified so that all the values are greater than zero.

Note that these populations were constructed under various scenarios for the relationship between the x and y variables. *Ppgamma20* and *ppgamma5a* satisfy the assumptions of the super population model leading to *estcd*, while *ppgamma5b* is consistent with the assumptions underlying *estpp*. In *ppstskew* the conditional variance of y_i given x_i is consistent with *estcd* while for the unmodified *ppln* it was consistent with *estpp*. In both these cases the assumption for the conditional expectation is not satisfied. For the populations *ppcounties*, *ppgamma5a* and *ppln* we have plotted y against x and y/x against x. The results are seen in Figures 1 through 3.

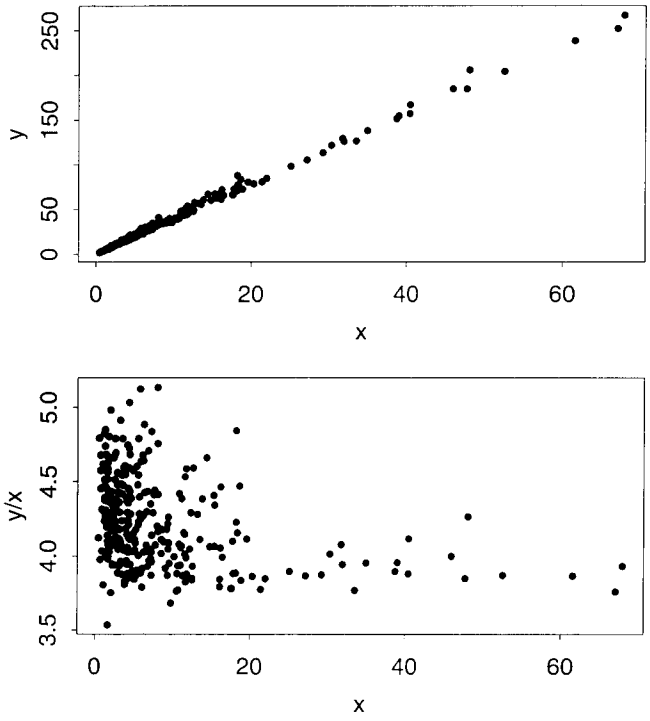

Figure 1. For *ppcounties* the plot of y versus x and y/x versus x where x is the number of families (thousands) living in a county and y is the total population (thousands) of the county for 304 counties.

The estimator *estpp* is based on the assumption that given the sample s our beliefs about the observed ratios, *i.e.*, the ratios y_i/x_i for $i \in s$ and the unobserved ratios, *i.e.*, the ratios y_j/x_j for $j \notin s$ are roughly exchangeable. In particular this means that one's beliefs about a ratio y_j/x_j

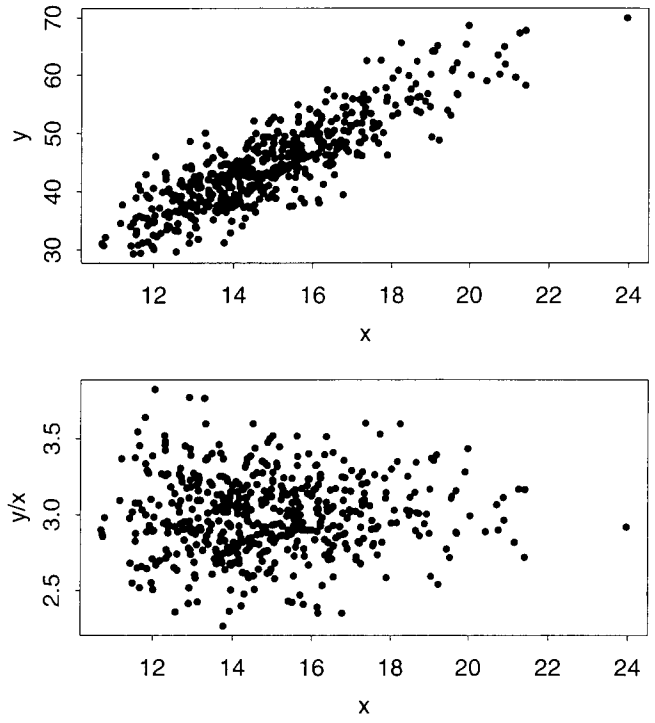

Figure 2. For *ppgamma5a* the plot of y versus x and of y/x versus x.

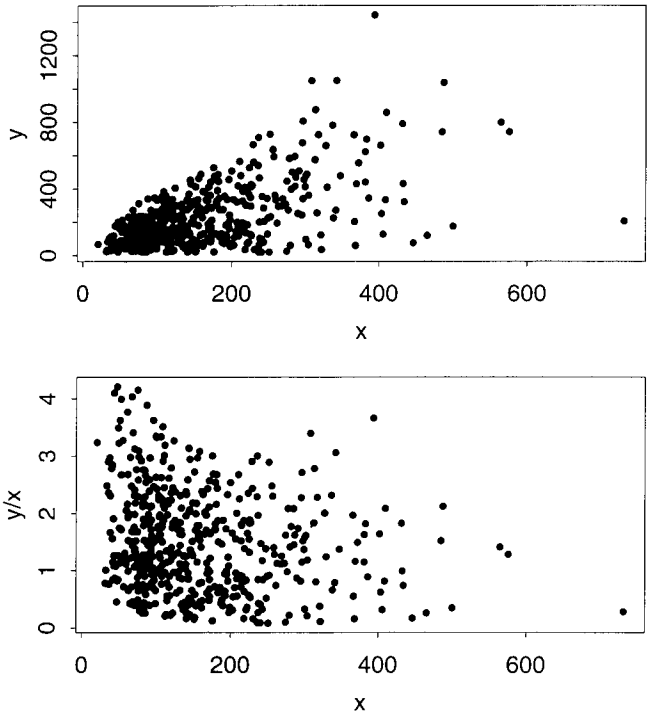

Figure 3. For *ppln* the plot of y versus x and of y/x versus x.

should not depend on the size of x_j. In fact *ppgamma5b* was constructed so that this would indeed be true. On the other hand, under the super population model leading to the estimator *estcd* we would expect the variability of the ratios to get smaller as the size of the x variable increases while the average value of the ratios in any thin vertical strip remains roughly constant as the strip moves to the right. This is seen clearly in the plot of the ratios for population *ppgamma5a*. For the rest of the populations, except for *ppgamma20* the values of the ratios do in fact depend on size of x. This is seen clearly in the plots for *ppcounties* and *ppln*. Hence they should make interesting test cases for the estimator *estpp*. *Ppexp* was included as a test case to see what would happen if the underlying assumptions of *estpp* and *estcd* were strongly violated.

4. SOME SIMULATION RESULTS

To compare the six estimators 500 simple random samples of various sizes were taken from the nine populations. For each sample the values of the six estimators were computed. For the estimator *estpp* this meant finding it approximately by simulating $R = 500$ realizations of the predictive distribution for the population median induced by the 'Polya posterior'. In each case the average value and average absolute error of the estimator were computed. In Table 1 the average values of all the estimators except *estsm* are given. All the estimators are approximately unbiased except in one case, *estcd* for the population *ppln*. We did not include the results for *estsm* since it is well known that it is unbiased. In Table 2 the average absolute error for all six estimators are given. We see from Table 2 that *estcd* and *estpp* are the clear winners. They both perform better than the other four estimators in every case but one. In *ppexp* they are both beaten by *estsm*, but this is one case where neither would be expected to do well. For the first seven populations their performances are nearly identical while for population *ppln* the estimator *estpp* is preferred and for population *ppstskew* the opposite is true.

In practice one often desires interval estimates as well as point estimates for parameters of interest. Kuk and Mak (1989) and Chambers and Dunstan (1986) each suggested possible methods for finding interval estimates based on their estimator using asymptotic theory. But in each case they did not actually find any interval estimators. Meeden and Vardeman (1991) noted how approximate 95% credible regions based on the 'Polya posterior' can be found approximately. If we let $q(.025)$ and $q(.975)$ be the .025 quantile and the .975 quantile of the collection of 500 simulated population medians under the 'Polya posterior' then $(q(.025), q(.975))$ is an approximate 95% credible interval. (See Berger 1985 for the definition of such intervals.) Table 3 gives the average length and relative frequency of coverage for these intervals. We see that for these populations the intervals have reasonable frequentist properties. Perhaps this is not unexpected given the discussion in Meeden and Vardeman (1991). But on the other hand only one of the populations was constructed so that the ratios y_i/x_i are exchangeable. These results suggest that point and interval estimators of the median based on the 'Polya posterior' for the ratios are fairly robust against the exchangeability assumption and should work well in a variety of situations. This will be discussed further in section 5.

Table 1

The Average Value of Five Estimators of the Median for 500 Simple Random Samples

Population (median)	Sample Size	Average Value of the Estimator				
		estrm	*estkm*	*estrkm*	*estcd*	*estpp*
ppcities (1.90)	25	.197	.196	.193	.195	.195
ppsales (1.24)	30	1.21	1.25	1.23	1.25	1.24
ppcounties (18.33)	30	18.21	18.60	18.66	18.26	18.39
ppexp (29.02)	30	29.03	29.05	29.00	29.03	29.05
ppgamma5a (43.90)	30	43.82	43.88	43.91	43.99	43.89
	50	43.90	43.91	43.85	44.06	43.90
ppgamma5b (44.17)	30	43.84	43.96	44.19	44.15	43.61
	50	44.28	44.37	44.18	44.18	43.98
ppgamma20 (23.15)	30	23.47	23.28	23.14	23.46	23.77
	50	23.34	23.18	23.17	23.43	23.18
ppln (170.25)	30	171.15	169.38	168.12	185.01	170.61
	50	169.15	167.54	167.65	185.03	169.61
ppstskew (46.12)	30	43.66	40.27	45.88	45.50	45.11
	50	44.04	40.70	46.01	45.43	45.37

Table 2

The Average Absolute Error of Six Estimators of the Median for 500 Simple Random Samples

Population	Sample Size	Average Absolute Error of the Estimator					
		estsm	*estrm*	*estkm*	*estrkm*	*estcd*	*estpp*
ppcities	25	.0326	.0161	.0162	.0155	.0075	.0072
ppsales	30	.1797	.0770	.0797	.0870	.0244	.0245
ppcounties	30	3.12	.586	.964	1.34	.215	.214
ppexp	30	.43	.49	.48	.47	.48	.46
ppgamma5a	30	1.36	.96	1.03	.89	.54	.53
	50	.95	.74	.78	.65	.44	.43
ppgamma5b	30	2.84	2.74	2.71	2.58	2.37	2.38
	50	2.08	2.04	2.01	1.89	1.80	1.85
ppgamma20	30	1.08	1.06	1.05	.88	.67	.64
	50	.94	.77	.78	.73	.51	.49
ppln	30	25.9	25.8	24.2	21.62	21.4	17.0
	50	18.0	20.1	17.9	16.46	17.7	12.7
ppstskew	30	3.86	4.26	6.69	3.21	2.72	3.14
	50	2.92	3.63	5.82	2.55	2.20	2.51

Table 3

The Average Length and Relative Frequency of Coverage for a .95 Credible Interval for the Median Based on the 'Polya Posterior' for 500 Simple Random Samples

Population	Sample Size	Average Length	Frequency of Coverage
ppcities	25	.041	.968
ppsales	30	.141	.964
ppcounties	30	1.44	.994
ppexp	30	2.26	.944
ppgamma5a	30	2.70	.950
	50	2.15	.956
ppgamma5b	30	11.67	.932
	50	8.86	.942
ppgamma20	30	3.24	.960
	50	2.51	.964
ppln	30	84.8	.934
	50	65.4	.956
ppstskew	30	15.52	.936
	50	12.00	.938

5. DISCUSSION

The motivation for the estimator *estpp* is based on the assumption that the population ratios y_i/x_i's are exchangeable. This assumption can be described mathematically in two separate but related ways. The first is the super population model given earlier while the second comes from the 'Polya posterior' which is based on a stepwise Bayes argument and gives a non-informative Bayesian interpretation for the estimator. This second approach can be used no matter what parameter is being estimated. When estimating the mean it leads to Basu's estimator which performs very much like the ratio estimator although the ratio estimator usually does a bit better. When estimating the median it leads to the estimator discussed in this note. Here we have argued that the 'Polya posterior' for the ratios leads to good point and interval estimators for the median when an auxiliary variable is present and seems to be reasonably robust against the assumption that the ratios y_i/x_i's are exchangeable.

Royall and Cumberland (1981) gave an empirical study of the ratio estimator and estimators of its variance. They argued that given a sample an estimate of variance based on the super population model, which leads to the ratio estimator, often made more sense than a design based estimate based on a probability sampling distribution. In Royall and Cumberland (1985), they demonstrated that, conditional on the sample mean of the auxiliary variable, the conditional coverage properties of the usual designed based confidence interval for the population mean were 'hopelessly unreliable'.

We now wish to address the question of the conditional behavior of the intervals for the median based on the Polya posterior which were developed in this note. In the simulation studies given earlier simple random sampling was used for convenience. To get some idea of the conditional behavior of the 'Polya posterior' we considered five of our populations. In each case we ordered the population using the values of the auxiliary variable x. We then took 500 random samples from the first or smallest half of the population, then 500 more random samples from the second or largest half of the population and finally 500 more random samples from the middle third of the population. We then calculated the .95 credible interval for the median based on the 'Polya posterior' which assumes the exchangeability of the ratios y_i/x_i's. In Table 4 we give the results for the 'Polya posterior' estimators for the median. (We also computed the average value and average absolute error of *estcd* for these examples. We did not include these results since they match closely the results of the 'Polya posterior'.) We see that their conditional behavior, at least in these cases, is very much like their unconditional behavior. In short, interval estimates for the median based on the 'Polya posterior' should have reasonable frequentist properties, no matter how the sample was selected, as long the population approximates our beliefs that the ratios are roughly exchangeable.

Table 4

The Average Value and Absolute Error for the Point Estimator and the Average Length and Relative Frequency of Coverage for a .95 Credible Interval for the Median Based on the 'Polya Posterior' for 500 Simple Random Samples from the whole Population, the 'Smallest' Half, the 'Largest' Half and the 'Middle' Third

Population	Sample Size	Where Taken	Average Value	Average Error	Average Length	Frequency of Coverage
ppcities	25	whole	.195	.0072	.041	.968
		smallest ½	.192	.0047	.033	.994
		largest ½	.196	.0078	.048	.988
		middle ⅓	.201	.0114	.055	.922
ppcounties	30	whole	19.4	.220	1.46	.990
		smallest ½	18.6	.305	1.34	.942
		largest ½	18.1	.283	1.59	.954
		middle ⅓	18.5	.252	1.35	.964
ppsales	30	whole	1.24	.0072	.141	.964
		smallest ½	1.24	.027	.153	.966
		largest ½	1.23	.020	.125	.982
		middle ⅓	1.23	.027	.139	.944
ppgamma5a	30	whole	43.9	.53	2.70	.950
		smallest ½	43.8	.55	2.82	.948
		largest ½	44.0	.53	2.55	.940
		middle ⅓	43.9	.47	2.63	.974
ppgamma5b	30	whole	43.6	2.38	11.7	.932
		smallest ½	42.2	2.69	11.6	.890
		largest ½	45.1	2.25	11.2	.950
		middle ⅓	45.2	2.27	11.3	.936

As can be seen by looking at the plots of y_i/x_i versus x_i and our simulation results it does not seem to matter much if the variability in the ratios y_i/x_i's decreases as x_i increases. What is crucial however is that the average value of the ratios in the narrow strip above a small interval of possible x values remains fairly constant as we move the small interval to the right. In Figure 2, the plot of the ratios for *ppgamma5a* is an example of such a plot. In fact this is how the population was constructed, since it satisfies the assumptions underlying *estcd*. In Figures 1 and 3 we see for *ppcounties* and *ppln* that the average value of the ratios in a narrow strip tends to decrease as we move to the right and helps to explain the relatively poorer performance of the 'Polya posterior' estimators in these cases. Overall however, the performance of procedures based on the 'Polya posterior' seem to be reasonably robust against the exchangeability assumption.

As another alternative we could consider a more balanced sampling plan which is based on stratifying the population on the auxiliary variable. For example consider again population *ppgamma5b* where it is ordered on the basis of its x_i values. We constructed ten strata where the first stratum consisted of the units with the fifty smallest x_i values, the second stratum of the units with the next fifty smallest x_i values and so on. We then took 500 stratified random samples of size fifty where five units were chosen at random from each stratum. For these samples the average value of *estpp* was 43.94 and its average absolute error was 1.81. The average length of its corresponding interval estimator was 8.95 with .938 relative frequency of covering the true value. Note that these figures are very similar to those given Tables 1 and 2 when simple random sampling was used.

ACKNOWLEDGEMENTS

Research supported in part by NSF grant SES 9201718.

REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling, part one. In *Foundations of Statistical Inference*. Toronto: Holt, Reinhart and Winston, 203-242.

BERGER, J.O. (1985). *Statistical Decision and Bayesian Analysis*. New York: Springer-Verlag.

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

CHAMBERS, R.L., and DUNSTAN, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

FERGUSON, T.S. (1973). A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1, 209-230.

KUK, A.Y.C., and MAK, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, 51, 261-269.

MAK, T.K., and KUK, A. (1993). A new method for estimating finite-population quantiles using auxiliary information. *The Canadian Journal of Statistics*, 21, 29-38.

MEEDEN, G., and GHOSH, M. (1983). Choosing between experiments: applications to finite population sampling. *Annals of Statistics*, 11, 296-305.

MEEDEN, G., and VARDEMAN, S. (1991). A noninformative Bayesian approach to interval estimation in finite population sampling. *Journal of the American Statistical Association*, 86, 972-980.

MEEDEN, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics and Probability Letters*, 16, 103-109.

RAO, J.N.K., KOVAR, J.G., and MANTEL, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

ROYALL, R.M., and CUMBERLAND, W.D. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

ROYALL, R.M., and CUMBERLAND, W.D. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.