A decision theoretic approach to Imputation in finite population sampling

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

August 1997
Revised May and November 1999
To appear in JASA

*Research supported in part by NSF Grant DMS 9401191

A decision theoretic approach to Imputation in finite population sampling

SUMMARY

Consider the situation where observations are missing at random from a simple random sample drawn from a finite population. In certain cases it is of interest to create a full set of sample values such that inferences based on the full set will have the stated frequentist properties even though the statistician making those inferences is unaware that some of the observations were missing in the original sample. We will give a Bayesian decision theoretic solution to this problem when one is primarily interested in making inferences about the population mean.

Key Words: finite population sampling, imputation, missing values, decision theory.

AMS 1991 Subject Classification: Primary 62D05; Secondary 62C10.

1 Introduction

The problem of nonresponse is an important one in sample surveys and is difficult to handle in the usual frequentist formulation. Various suggestions have been made for imputing values for the missing observations but without much theoretical justification. The difficulty with most approaches that just replace each missing value with a single imputed value is that inferences based on the completed data set over estimate the precision of the inferential procedure. Handling nonresponse is theoretically more straight forward from the Bayesian point of view but in practice it also faces various difficulties.

An interesting approach to this problem is multiple imputation which is described in Rubin (1987). Multiple imputation was developed to handle missing data in public-use files where the user has only complete-data methods available and has limited information about the reasons for nonresponse. It uses Bayesian ideas to yield inference procedures with sensible frequentist properties. Rubin distinguishes between ignorable and nonignorable nonresponse. When the nonresponse is nonignorable and follow-up surveys are not possible, Rubin argues, correctly we believe, that any sensible analysis must be based on an assumed model for the nonresponse. In either case given a sample which contains some nonrespondents he constructs a distribution for the missing observations. A complete data set is then formed, by using this distribution to impute values for all the missing observations. Then the completed data set is analyzed using standard procedures just as if the imputed data were the real data obtained from the nonrespondents. This process is repeated several times where in each repetition a new set of imputed values is chosen for the missing observations. This collection of complete-data inferences can be combined to form one inference that more properly reflects the uncertainty due to nonresponse than is possible if just one set of imputed values is considered.

In what follows we will always assume that the observations are missing at random, i.e. that the nonresponse is ignorable. In addition we must assume that the variable of interest is continuous. The methods presented here will not work with categorical or discontinuous variables which are often present in survey data. We assume that what is wanted is a single full data set that can be analyzed by a variety of users who will either just ignore or be ignorant of the fact that some of the original observations were missing. We will mainly be interested in the problem of making inferences about the population mean. It is intuitively clear that if one just imputes the miss-

ing values in the sample without considering the variance of the completed sample then the measures of precision based on this completed sample will almost certainly be wrong. Recently Rao and Shao (1992) considered a procedure that constructs a single completed sample but then used a jackknife type estimate of variance which treats the imputed values differently than the observed values to get a "correct" estimate of variance. For a recent discussion of multiple imputation and an extension of the work of Rao and Shao see Rubin (1996), Fay (1996) and Rao (1996) and the related discussion. In another context Nusser et al. (1996) recently suggested replacing data from a complex survey with unequal weights by a created sample whose members could be given equal weights.

This suggests when the sample contains nonresponders that in addition to imputing values for the missing observations to get a completed sample one could then adjust either just the imputed values or all the values in the completed sample so that the resulting created sample has the appropriate mean and variance. In section 2 we will give a decision theoretic formulation for the problem of finding a created sample that can be used in place of a sample with missing values. The Bayesian solution is found for the simplest situation where the characteristic of interest is univariate and there are no auxiliary variables. In section 3 this approach is then extended to the situation where an auxiliary variable is present and a noninformative stepwise Bayesian solution is found. In section 4 created samples are found which yield correct inferences when the ratio estimator will be used and observations are missing. In section 5 we extend the results of section 2 to the situation where an additional constraint on the range of the created sample has been added. In section 6 we apply this approach to the problem where each unit in the population has associated with it a vector of variables of interest some of which may be missing and we wish to construct a created sample with no missing values. In Section 7 we will give a brief discussion of the merits of created samples.

2 A decision problem

We begin by introducing the standard notation of finite population sampling for the simplest situation. Let $\mathcal{U}$ denote a finite population which consists of N units labeled $1, 2, \ldots, N$. Attached to unit i let y_i, a real number, be the unknown value of some characteristic of interest. For this problem

$\mathbf{y} = (y_1, \ldots, y_N)$ is the unknown state of nature or parameter. Each $\mathbf{y}$ is assumed to belong to $\mathcal{Y}$ some subset of N-dimensional Euclidean space. A subset s of $\{1, 2, \ldots, N\}$ is called a sample. Let $n(s)$ denote the number of elements belonging to s. Let S denote the set of all possible samples. A (nonsequential) sampling design is a function p defined on S such that $p(s) \in [0, 1]$ for every nonempty $s \in S$ and $\sum_{s \in S} p(s) = 1$. Given $\mathbf{y} \in \mathcal{Y}$ and $s = \{i_1, \ldots, i_{n(s)}\}$, where $1 \leq i_1 < \cdots < i_{n(s)} \leq N$ let $\mathbf{y}(s) = (y_{i_1}, \ldots, y_{i_{n(s)}})$. Since we are assuming that nonresponse can be present $\mathbf{y}(s)$ may not be completely known to us. Let s_r denote the subset of s which contains the labels of the units in the sample for which y was actually observed. Hence $\mathbf{y}(s_r)$ is what we actually get to see. Let n be a fixed positive integer such that $1 < n < N$ and for simplicity we will assume that the design p is such that $p(s) = 0$ whenever $n(s) \neq n$. Let $n(s_r) = n_r$ denote the size of the set of responders in the sample. Let $\bar{\mathbf{y}}(s_r)$ and $\mathrm{Var}(\mathbf{y}(s_r))$ denote the sample mean and sample variance of the responders. For a fixed design and ignorable nonresponse mechanism (Rubin, 1987) we have now defined the sample space for our experiment.

The next step is to define the set of possible decisions or actions and the loss function. The set of actions will just be n-dimensional Euclidean space, $\mathcal{R}^n$, the set of all full samples of size n. Note that we are being imprecise here. More properly the set of all possible actions is the set of all created samples of size n for all possible samples s with $n(s) = n$. That is, for notational convenience we are suppressing the dependence of the action on the labels that appear in s. We will denote a typical action by $\mathbf{a} = (a_1, \ldots, a_n)$. Let $\bar{\mathbf{a}}$ and $\mathrm{Var}(\mathbf{a})$ denote the sample mean and sample variance of the sample which just consists of the values making up $\mathbf{a}$. Note the divisor in the sample variance is $n - 1$. We can now define the loss function. The loss will depend on $\mathbf{y}$, $\mathbf{a}$, s and s_r and is given by

$$L(\mathbf{a}, \mathbf{y}, s, s_r) = \frac{\sum_{j=1}^{n} (y_{i_j} - a_j)^2}{I(\bar{\mathbf{y}}(s_r) = \bar{\mathbf{a}}, \frac{n(N-n_r)}{n_r(N-n)} \mathrm{Var}(\mathbf{y}(s_r)) = \mathrm{Var}(\mathbf{a}))} \tag{2.1}$$

where I denotes the usual indicator function. Note that finding $\mathbf{a}$ to minimize this loss function is equivalent to finding the $\mathbf{a}$ which minimizes $\sum_{j=1}^{n} (y_{i_j} - a_j)^2$ subject to the constraints

$$\bar{\mathbf{a}} = \bar{\mathbf{y}}(s_r) \quad \text{and} \quad \mathrm{Var}(\mathbf{a}) = \frac{n(N - n_r)}{n_r(N - n)} \mathrm{Var}(\mathbf{y}(s_r)). \tag{2.2}$$

Assume the design is just simple random sampling without replacement of size n and suppose that if there were no missing observations we would use the usual normal theory interval to construct a 95% confidence interval for the population mean. Assuming the nonresponse is ignorable and the sample contains missing observations we could just use the set of responders to create correct frequentist intervals. However in the scenario described above it is of interest to create a new "sample" of size n so that a data analyst who is unaware that any observations were missing could use this created sample and the usual normal theory to construct a 95% confidence interval and the procedure will have the correct frequentist properties under repeated sampling. Note that any created sample of size n whose mean agrees with $\bar{\mathbf{y}}(s_r)$ and whose variance is such that when used to compute the standard interval estimate it gives the same length interval as the standard interval based just on the set of responders will do the trick. The loss function defined above was chosen to select such a created sample in a sensible way. Now the numerator of the loss function is just the squared distance between $\mathbf{y}(s)$ and the created sample. Hence for a given $\mathbf{y}$ and a given sample s this loss function selects the created sample which satisfies the two constraints (which depend just on $\mathbf{y}(s_r)$, the values of the responders in the sample) and is closest to $\mathbf{y}(s)$.

Our next step is to find the Bayes solution for this decision problem. Let π denote a probability density function defined on the parameter space $\mathcal{Y}$. Now it is well known that in this setup the conditional distribution of the nonresponders in the sample and the rest of the unsampled units in the population given the values of the responders in the sample, $\mathbf{y}(s_r)$, is just the appropriate conditional distribution found in the usual way and does not depend on the design probabilities. Hence we can also find the conditional distribution of the nonresponders in the sample by integrating over the unsampled units in the previous conditional distribution.

From now on for notational convenience we will always assume that the sample s is just the first n units of the population and that the responders are just the first n_r elements of the sample. For $j = n_r + 1, \ldots, n$ let y_j^* denote the posterior expectation of y_j for the nonresponder j. Now since under these assumptions

$$\sum_{j \in s} (y_{i_j} - a_j)^2 = \sum_{j=1}^{n_r} (y_j - a_j)^2 + \sum_{j=n_r+1}^{n} (y_j - a_j)^2$$

we see that finding the created sample $\mathbf{a}$ which minimizes the posterior expected loss of the loss function given in equation 2.1 is equivalent to finding the created sample $\mathbf{a}$ which minimizes

$$\sum_{j=1}^{n_r}(y_j - a_j)^2 + \sum_{j=n_r+1}^{n}(y_j^* - a_j)^2$$

subject to the mean and variance of $\mathbf{a}$ satisfying the constraints of equation 2.2 which depend on $\mathbf{y}(s_r)$. This means that we are trying to find the created sample which is closest to the vector which consists of the values of the responders and our best guess for the nonresponders and which satisfies the appropriate constraints. In the following theorem we show that the solution for this problem is easily found using Lagrange multipliers. For a similar result see Ghosh (1992).

Theorem 2.1 *Let $n \geq 2$ be a fixed positive integer and let $\mathbf{y}^*$ be a fixed vector of real numbers of length n. Let $\bar{\mathbf{y}}^*$ and $\mathrm{Var}(\mathbf{y}^*)$ denote the sample mean and sample variance of the "sample" which just consists of the values making up $\mathbf{y}^*$.*

Let m and $u > 0$ be fixed real numbers. Let $\mathbf{a}$ be a vector of real numbers of length n. For the problem of minimizing $\sum_{i=1}^{n}(a_i - y_i^*)^2$ over $\mathbf{a}$ subject to the constraints that $\bar{\mathbf{a}} = m$ and $\mathrm{Var}(\mathbf{a}) = u$ the solution is given by $\mathbf{a}^o$ where

$$a_i^o = m + \left(\frac{u}{\mathrm{Var}(\mathbf{y}^*)}\right)^{1/2}(y_i^* - \bar{\mathbf{y}}^*).$$

Proof. Using the method of Lagrange multipliers we consider the unconstrained problem of minimizing

$$\sum_{i=1}^{n}(a_i - y_i^*)^2 - 2\gamma_1\sum_{i=1}^{n}a_i - \gamma_2\sum_{i=1}^{n}(a_i - m)^2$$

Taking the partial derivative of the above equation with respect to a_i and setting it equal to zero we get

$$a_i = \frac{y_i^* + \gamma_1 - \gamma_2\, m}{1 - \gamma_2}$$

Now summing the above equation over i and remembering that the first constraint is $\bar{\mathbf{a}} = m$ we find that

$$\gamma_1 = m - \bar{\mathbf{y}}^*$$

and substituting this into the previous equation we have that

$$a_i = m + \frac{y_i^* - \bar{y}^*}{1 - \gamma_2}$$

Next we use the above expression for a_i and both of the constraints to find

$$u = \frac{\sum_{i=1}^n (a_i - m)^2}{n-1} = \frac{\sum_{i=1}^n (y_i^* - \bar{y}^*)^2}{(n-1)} \frac{1}{(1-\gamma_2)^2}$$

or

$$\frac{1}{1-\gamma_2} = \left(\frac{u}{\mathrm{Var}(\mathbf{y}^*)} \right)^{1/2}$$

and the expression for the minimizing $\mathbf{a}^o$ given in the statement of the theorem follows. We took the positive square root for the solution for $1/(1-\gamma_2)$ in the above since the negative square root gives the solution to the problem of finding the $\mathbf{a}$ which is furthest away from $\mathbf{y}^*$ and satisfies the constraints. This completes the proof.

To apply the theorem to our sampling problem recall that for a given sample $\mathbf{y}^*$ will be the vector of values for the responders in the sample along with our posterior expectation of the nonresponders in the sample. Also m will be $\bar{y}(s_r)$, the mean of the responders in the sample, and u will be the function of the variance of the responders in the sample given in equation 2.2. Our goal is to find a created sample as close to $\mathbf{y}^*$ as possible and still have the two constraints satisfied. Note that the solution $\mathbf{a}^o$ is intuitively sensible. For a given i it begins by setting a_i^o equal to m the value of the first constraint and then adjusts it upward or downward by a fixed multiple of $y_i^* - \bar{y}^*$ where the direction depends upon the sign of this last term and the multiple is essentially a ratio of standard deviations.

If the parameter space $\mathcal{Y}$ is N-dimensional Euclidean and the prior density π is strictly positive on $\mathcal{Y}$ then the theorem gives the unique Bayes estimator with respect to π for any sample s with missing observations for constructing a created sample that is close to the observed data. This uniqueness implies that the Bayes solution is admissible for this problem. For a particular sample the values in the created sample will depend on both the observed data and the prior π. However frequentist properties of a confidence interval for the mean based on this procedure will be correct since the created sample will always have the correct mean and variance.

We conclude this section with another problem that can be solved by the first theorem. Nusser et al. (1996) considered data from a complex survey that resulted in samples with unequal weights. (We are assuming here that there are no missing values in the sample.) In the course of their study it became convenient to replace such a sample by an equal-weight sample. To create such an equal-weight sample that was close to the original unequal-weight sample they began by constructing the empirical cumulative distribution function for the unequal-weight sample. Next they created a continuous distribution function which approximated the empirical cumulative distribution function by connecting the midpoints of the rises in the steps of the empirical cumulative distribution function in a piecewise linear fashion. Finally they took the inverse of the continuous approximation at a set of equally spaced quantiles to get the created equal-weight sample.

Let $\mathbf{y}^*$ denote the observed unequal-weight sample of length n. We assume that the weights are strictly positive and sum to one. Let m and $u > 0$ denote the the mean and variance of $\mathbf{y}^*$ under the probability distribution given by the weights. Now what we want to find is a vector $\mathbf{a}$ whose sample mean and sample variance, i.e. we are assuming equal weights, are equal to m and u respectively and whose values are close to $\mathbf{a}$. That is we want to find the vector $\mathbf{a}$ which minimizes $\sum_{i=1}^{n}(a_i - y_i^*)^2$ subject to constraints on the mean and the variance which is just the problem we have been considering.

3 Noninformative Bayesian imputation

In the previous section we assumed a Bayesian model to generate imputed values for missing observations. However the theorem can be applied to the completed vector no matter how the imputed values were selected. In this section we consider what can be done if there is little prior information available about the population. Meeden and Vardeman (1991) gave a noninformative Bayesian approach to interval estimation in finite population sampling. For further details see Ghosh and Meeden (1997). This approach is appropriate when given the observed values in the sample the statistician's beliefs about the seen and the unseen are roughly exchangeable. Given the responders in the sample inferences are based on a pseudo posterior called the Polya posterior. This conditional distribution for the unseen and in particular for the nonresponders in the sample is just usual Polya sampling from an urn which contains the responders in the sample. See Meeden and Bryan

(1996) for further details. This approach has a stepwise Bayes justification and one can show that procedures based on the Polya posterior are typically admissible. This argument is easily adapted to prove admissibility for the problem considered here. Since we are assuming ignorable nonresponse given the responders in the sample the conditional expectation of a typical nonresponder is just $\bar{\mathbf{y}}(s_r)$ the mean of the responders. Hence give a sample s, assumed to be the first n units in the population, the vector $\mathbf{y}^*$ to be used in the theorem is

$$y_i^* = \begin{cases} y_i & \text{for } i = 1, \ldots, n_r \\ \bar{\mathbf{y}}(s_r) & \text{for } i = n_r + 1, \ldots, n. \end{cases}$$

Now since $\mathrm{Var}(\mathbf{y}^*) = \{(n_r - 1)/(n-1)\}\,\mathrm{Var}(\mathbf{y}(s_r))$ we see from the theorem that the created sample $\mathbf{a}^o$ which is the solution for our problem is given by

$$a_i^o = \begin{cases} \bar{\mathbf{y}}(s_r) + \left(\frac{(n-1)n(N-n_r)}{(n_r-1)n_r(N-n)}\right)^{1/2}\left(y_i - \bar{\mathbf{y}}(s_r)\right) & \text{for } i = 1, \ldots, n_r \\ \bar{\mathbf{y}}(s_r) & \text{for } i = n_r + 1, \ldots, n. \end{cases}$$

Note that this is similar to mean imputation that just replaces each nonresponder by the mean of the responders. However here we have adjusted the responders in the sample as well. Note that the factor which we multiply $y_i - \bar{\mathbf{y}}(s_r)$ for each responder must be greater than one. Hence each responder is adjusted further away from the mean of the responders. This spreads the sample out. Intuitively something like this must be done if we want to replace the nonresponders by a fixed value and still get a sensible measure of variability.

This approach can also be used in stratified sampling when one is interested in making inferences about the population mean. For a stratum with missing observations a created sample can be found with the proper mean and variance. Since the estimated variance of the usual estimator of the mean is just a weighted average of the strata variances this will guarantee a procedure with the stated frequentist properties under repeated sampling.

To see how our method works in practice we consider an extension of the Polya posterior which incorporates prior information in the form of an auxiliary variable $\mathbf{x} = (x_1, \ldots, x_N)$. We assume that $y_i > 0$ and $x_i > 0$ for $i = 1, \ldots, N$. We will assume that the statistician's beliefs about the ratios $r_i = y_i/x_i$'s are roughly exchangeable. Such an assumption could be reasonable if the x_i's do not differ too much in size. Following Basu (1971) and Royall (1970) and assuming the nonresponse is ignorable, after the data

$\mathbf{y}(s_r)$ is observed, $\bar{\lambda}_{s_r} = n_r^{-1} \sum_{i=1}^{n_r} (y_i/x_i)$ is a sensible estimate of y_j/x_j for a nonresponding unit j. That is $\bar{\lambda}_{s_r} x_j$ is a sensible estimate of y_j and this is the value we will use in $\mathbf{y}^*$ for nonresponder j. This also has a stepwise Bayes justification where the Polya posterior is applied to the ratios y_i/x_i (Meeden and Ghosh, 1983).

Rather than selecting a created sample in which both the imputed values and the responders are adjusted one might wish to keep the responders fixed and just adjust the imputed values for the nonresponders in such a way that one gets the correct mean and variance for the created sample containing both sets of values. This problem can be solved in exactly the same way except that now $\mathbf{y}^*$ of the theorem will just contain the imputed values for the nonresponders. Clearly the constraint for the mean for the solution vector $\mathbf{a}^o$ must be $\bar{\mathbf{y}}(s_r)$ however the constraint for the variance must be adjusted so that the variance for the final created sample is correct. Once this is done the theorem again yields the solution except in the special case where each element of $\mathbf{y}^*$ is the same. For example consider the case where each element of $\mathbf{y}^*$ is just $\bar{\mathbf{y}}(s_r)$. Recall this is what happens under the Polya posterior. Without loss of generality we can assume that the mean of the responders is zero. In that case we want to minimize $\sum_{i=1}^{n-n_r} a_i^2$ subject to $\bar{\mathbf{a}} = 0$ and $\sum_{i=1}^{n-n_r} a_i^2 = u$ for some positive number u. The set of solutions is just where the hyperplane defined by the first constraint intersects with the surface of the sphere defined by the second constraint and is no longer given by the expression in the theorem since now $\mathrm{Var}(\mathbf{y}^*)$ is zero.

The created samples have been selected with the goal of making inferences about the population mean. In general we can not expect them to do well for all possible inferential problems. However it is possible that they might perform acceptably for some other problems. To this end we consider the problem of finding an interval estimator for the population median. We will use the asymptotic version of the interval estimate due to Woodruff (Särndal et al., 1992).

We consider an artificial population with an auxiliary variable, $\mathbf{x}$, of size 500 which was created as follows. Each x_i was set equal to ten plus an observation from a gamma distribution with shape parameter five and scale parameter one. Then y_i was set equal to an observation from a normal distribution with mean $3x_i$ and variance x_i. All the above observations were independent. The median of the population is 43.90 and the correlation between the y_i's and x_i's is 0.87. The superpopulation model under which this population was constructed gives the ratio estimator when estimating the

population mean. The values of the ratios y_i/x_i's do not depend on the x_i's but their variability decreases as the value of x_i increases. Hence this population does not satisfy the assumption that one's beliefs about the ratios are exchangeable. However it was noted in Meeden (1995) that when estimating the population median and using the auxiliary variable the stepwise Bayes model described above which uses the observed ratios in the sample and assumes that $\mathbf{x}$ is completely known yields a sensible interval estimator of the population median. Here we will assume that only the x_i's in the sample are known, i.e. we learn the value of x_i in the sample whether they respond or not. After the data $\mathbf{y}(s_r)$ is observed, we will use $\bar{\lambda}_{s_r} x_j$ as the imputed value of y_j for nonresponder j where $\bar{\lambda}_{s_r}$ is the average of the ratios of the responders.

Some simulation results are given in Table 1. Given a simpe random sample of size fifty we considered four possible samples. The first was just the full sample, denoted by full in the table. For the full sample we calculated its range, median and the absolute error of its median as an estimate of the population median. We also computed the Woodruff 95% confidence interval for the population median. We noted its lower bound and length and noted if it contained the population median. Next we assumed that only a fixed number of units will respond say n_r and selected n_r responders from the full sample to get our sample of responders which we denote by resp. We then calculated the same seven quantities for resp that we did for full. Note that the values of the auxiliary variable so far have played no role. Next we used the values of the auxiliary variable for all the units in the full sample to construct imputed values for all the nonresponders in the sample. We then construct two new samples using the theorem and the imputed values. In the first we allowed both the responders' values and the imputed values to be adjust to get our created sample. In the second we held the responders' values fixed and only allowed the imputed values be adjusted to get our created sample. The first created sample we denoted by adall while the second was denoted by adimp. For these two samples we then computed the same seven quantities as we did for the earlier two samples. We did this for 500 simple random samples of size 50 for three different choices of n_r. The values of n_r were 45, 35 and 25. We only included the results for the full sample in the $n_r = 45$ case since the other cases are similar.

The first thing we see in Table 1 is that the range of the sample adall is larger than the range of the sample resp. Moreover the difference increases as the number of responders decreases. The same is true for sample adimp.

However the point estimator based on the sample adall has smaller absolute error on the average than the point estimator based on the sample resp and it gives shorter confidence intervals than those based on sample resp. Moreover the size of the improvement increases as the number of responders decreases. Perhaps this is not so surprising once we remember that sample adall is using the additional information that is contained in the auxiliary variable. Sample adimp is using this information as well but not as effectively as sample adall. Comparing sample full to sample adall for individual samples you find that sample adall has a few more extreme values and is less concentrated in the center than the other, much as you would expect. This suggests that sample adall should work quite well when estimating the population median even when the level of nonresponse gets quite high if there is enough information to make reasonable imputations for the nonresponders. We would not expect such good behavior if we wanted to make inferences about the 0.75 or 0.95 quantile and the nonresponse was more than minimal.

put table 1 about here

Up until now we have been assuming that the guessed values for the nonresponders were found under either a Bayesian Model or a stepwise Bayesian model. Formally this allowed us to prove admissibility for the decision problem formulated here. However one can apply the theorem once $\mathbf{y}^*$ has been selected no matter what imputation method was used. Hence any imputation scheme can be used to assign values to the nonresponders and then the theorem can be applied to find a created sample which will yield the correct frequentist mean and variance.

4 Imputation for the Ratio Estimator

Suppose that in addition to the characteristic y_i an auxiliary variable, x_i, closely related to y_i is observed for every unit in the sample. Also suppose that the population mean of the auxiliary variable, $\bar{X}$, is known. In this setting when estimating the population mean of $\mathbf{y}$ the ratio estimator is often the estimator of choice. It has been suggested that for a missing observation y_j a reasonable imputed value is $(\bar{\mathbf{y}}(s_r)/\bar{\mathbf{x}}(s_r))\bar{X}$. However the estimated variance of the ratio estimator using the completed sample which contains both the observed and imputed values will not be corrected. For more details

in the case of a stratified population see Rao(1996). There a "correct" estimate of the variance is found but the completed data set must also denote which of the y values have been imputed. Now we will show how to construct a created sample which will have the "correct" variance without needing any additional information.

If we did not desire a completed data set then given a sample with some missing y values we would just use the responders to estimate the population mean. The ratio estimate is just $\hat{R}_r\bar{X}$ where $\hat{R}_r = (\bar{\mathbf{y}}(s_r)/\bar{\mathbf{x}}(s_r))$ and $\bar{\mathbf{x}}(s_r)$ is just the mean of the auxiliary variable of the responders. Recall that its estimate variance is

$$\frac{1-f_r}{n_r(n_r-1)}\sum_{i=1}^{n_r}(y_i - \hat{R}_r x_i)^2$$

where $f_r = n_r/N$. Let $\mathbf{y}^*(s)$ be the responders augmented with imputed values for the nonresponders. The imputation could be done as suggested above but that plays no role in what follows. Now we want to find a created sample, $\tilde{\mathbf{y}}(s)$, which is close to $\mathbf{y}^*(s)$ and yields the same point estimate of the population mean as the one based just on the responders and has the "correct" estimate of variance.

The ratio point estimate of the mean based on $\tilde{\mathbf{y}}(s)$ is just $\tilde{R}\bar{X}$ where $\tilde{R} = \bar{\tilde{\mathbf{y}}}(s)/\bar{\mathbf{x}}(s)$. The two will agree when $\tilde{R} = \hat{R}_r$ or when

$$\bar{\tilde{\mathbf{y}}}(s) = (\bar{\mathbf{y}}(s_r)/\bar{\mathbf{x}}(s_r))\bar{\mathbf{x}}(s)$$

The estimates of variance will agree when

$$\sum_{i=1}^{n}(\tilde{y}_i - \hat{R}_r x_i)^2 = \frac{(1-f_r)n(n-1)}{(1-f)n_r(n_r-1)}\sum_{i=1}^{n_r}(y_i - \hat{R}_r x_i)^2$$

where $f = n/N$. So we want to find the created sample which is closest to $\mathbf{y}^*(s)$ and satisfies the above two constraints. So a solution will exist if the intersection of the above hyperplane and sphere are nonempty. The solution is found in the next theorem.

Theorem 4.1 *Let $n \geq 2$ be a fixed positive integer and let $\mathbf{y}^* \neq \mathbf{v}^*$ be fixed vectors of real numbers of length n. Let m and $u > 0$ be fixed real numbers which satisfy $|m - \bar{\mathbf{v}}^*| \leq \sqrt{u/n}$ where $\bar{\mathbf{v}}^*$ is the mean of the values making up $\mathbf{v}^*$. Let $\mathbf{a}$ be a vector of real numbers of length n. For the problem of*

minimizing $\sum_{i=1}^{n}(a_i - y_i^*)^2$ *over* $\mathbf{a}$ *subject to the constraints that* $\bar{\mathbf{a}} = m$ *and* $\sum_{i=1}^{n}(a_i - v_i^*)^2 = u$ *the solution is given by* $\mathbf{a}^o$ *where*

$$a_i^o = m - \bar{\mathbf{v}}^* + v_i + t_o(\bar{\mathbf{v}}^* - \bar{\mathbf{y}}^* + y_i - v_i)$$

and t_o *is one of the two values satisfying*

$$t_o = \pm\sqrt{\frac{u - n(m - \bar{\mathbf{v}}^*)^2}{\sum_{i=1}^{n}(\bar{\mathbf{v}}^* - \bar{\mathbf{y}}^* + y_i - v_i)^2}}$$

Proof. Since subject to the constraint $\sum_{i=1}^{n}(a_i - v_i^*)^2 = u$ the sum $\sum_{i=1}^{n} a_i$ is maximized for $a_i = v_i^* + \sqrt{u/n}$ and minimized for $a_i = v_i^* - \sqrt{u/n}$ we see that the condition $|m - \bar{\mathbf{v}}^*| \leq \sqrt{u/n}$ must be satisfied for this problem to have a solution. Hence geometrically our problem is to find the vector lying in the intersection of the sphere centered at $\mathbf{v}^*$ of radius $\sqrt{u}$ with the hyperplane $\sum_{i=1}^{n} a_i = nm$ which is closest to the vector $\mathbf{y}^*$.

The intersection of a n dimensional sphere with a hyperplane is an $n-1$ dimensional sphere whose center is on the hyperplane and is on the line passing through the center of the original n dimensional sphere and perpendicular to the hyperplane. So to find the solution to our problem the first step will be to find the center of this $n-1$ dimensional sphere. Next we will find the vector lying in the hyperplane which is closest to $\mathbf{y}^*$. The solution to our problem must be on the line joining this vector and the center of the $n-1$ dimensional sphere and on the sphere as well.

Since $1, \ldots, 1$ is a set of attitude numbers for the hyperplane $\sum_{i=1}^{n} a_i = nm$ the ith coordinate of the center of the $n-1$ dimensional sphere must be of the form $v_i^* + t$ for $i = 1, \ldots, n$. Remembering that this point must lie on the hyperplane we get that

$$\mathbf{w} = (v_1^* + m - \bar{\mathbf{v}}^*, \ldots, v_n^* + m - \bar{\mathbf{v}}^*)$$

is the center of the $n-1$ dimensional sphere.

Next consider the vector $\mathbf{z}$ where $z_i = y_i^* - \bar{\mathbf{y}}^* + m$ for $i = 1, \ldots, n$. It is easy to check that $\mathbf{z}$ lies on the hyperplane. Also since $y_i^* - z_i = m - \bar{\mathbf{y}}^*$ the direction numbers of the line joining $\mathbf{y}^*$ and $\mathbf{z}$ are proportional to the attitude numbers of the hyperplane. So this line is perpendicular to the hyperplane and $\mathbf{z}$ is the vector closest to $\mathbf{y}^*$ lying in the hyperplane. Now it is easy to check that the line joining the center of the $n-1$-dimensional sphere and $\mathbf{z}$ must be of the form

$$a_i = m - \bar{\mathbf{v}}^* + v_i^* + t(\bar{\mathbf{v}}^* - \bar{\mathbf{y}}^* + y_i - v_i^*)$$

Using the fact that the solution must lie on this line and on the sphere centered at $\mathbf{v}^*$ we can solve for two possible values of t one of which must yield the solution to our problem and the proof is complete.

So far we have been assuming that the created sample we select will adjust both the imputed values and the values of the responders. Suppose instead that we required only the imputed values be adjusted to get a solution. Then it is easy to check that our two constraints become

$$\bar{\tilde{\mathbf{y}}}(s_{nr}) = (\bar{\mathbf{y}}(s_r)/\bar{\mathbf{x}}(s_r))\bar{\mathbf{x}}(s_{nr})$$

where $\bar{\mathbf{x}}(s_{nr})$ is the mean of the auxiliary variable for the nonresponders and $\bar{\tilde{\mathbf{y}}}(s_{nr})$ is defined similarly and

$$\sum_{i=n_r+1}^{n} (\tilde{y}_i - \hat{R}_r x_i)^2 = \left(\frac{(1-f_r)n(n-1)}{(1-f)n_r(n_r-1)} - 1 \right) \sum_{i=1}^{n_r} (y_i - \hat{R}_r x_i)^2$$

Note that these two constraints for the adjustment of only the imputed values are exactly analogous to the two earlier constraints for adjusting the entire vector $\mathbf{y}^*(s)$. Now if we use as the imputed value for each missing y_j the value $\hat{R}_r x_j$ then in the theorem the center of the sphere $\mathbf{v}^*$ is equal to $\mathbf{y}^*$ the vector we are trying to approximate. In this case the theorem no longer applies but any point lying on the the intersection of the hyperplane and the sphere is a solution to our problem. For this reason in the example that follows we will not use $\hat{R}_r x_j$ as the imputed value when y_j is missing.

Consider again the artificial population used in the last section. Recall that to generate the y values for this population for a give value of x_i we set y_i equal to an observation from a normal population with mean $3x_i$ and variance x_i. This is a superpopulation model which yields the ratio estimator when estimating the mean. In practice one assumes that the variance of y_i given x_i is of the form $\sigma^2 x_i$ where σ is unknown. In this case given the responders in the sample a naive estimator of σ^2 is just

$$\hat{\sigma}^2 = \frac{1}{n_r - 1} \sum_{i=1}^{n_r} \left(\frac{y_i - \hat{R}_r x_i}{\sqrt{x_i}} \right)^2$$

This suggests the following naive imputation scheme. For a nonresponder with the value x_j for the auxiliary variable set its imputed y_j value equal to an observation from a normal distribution with mean $\hat{R}_r x_j$ and variance

$\hat{\sigma}^2 x_j$. Then one can use the theorem to find the created sample closest to this completed sample by either adjusting all the values in the completed sample or just adjusting the imputed values. In either case the created samples will yield inferences for the mean identical to those using just the responders when the ratio estimator is used.

To see how such created samples compare to the responders we took 500 random samples of size 50. For each sample we randomly selected 40 responders. We then used the imputation scheme described just above to form a completed sample. Finally we constructed two created samples where in one all the values in the completed sample were adjusted while in the other just the imputed values were adjusted. We denote the responders and these two created samples by resp, adall and adimp respectively. For each of these three samples we found their minimum values, their 0.25, 0.50 and 0.75 quantiles, their maximum values and their correlation with the auxiliary variable. The results are given in Table 2. Note that the samples seem quite similar although the ranges of adall and adimp are a bit longer than the range of resp and their correlation is a bit smaller.

Finally it should be noted that the results of this section extend easily to the stratified problem considered in Rao (1996) because that problem just reduces to constructing the correct created samples within each strata

put table 2 about here

5 Adding a Range Constraint

The complete sample would always be preferred to a created sample constructed from some of its values. Even though the created sample was selected to mimic important characteristics of the sample it could distort other other characteristics of interest. As we say in Table 1 the range of the created samples tends to be longer than the range of the completed samples. This suggests that one might also include a range constraint that the created sample must satisfy.

In general such a problem need not have a solution. This can happen when the range constraint is either too small or too large to be compatible with the variance constraint. Even if a solution does exist it can not be given the simple form of the solution for our earlier problem. The solution, when it exists, may have the property that several elements of the solution are

equal to to the minimum value of the solution and the same is true for the maximum value as well. We will not actually solve this general problem but a simpler one which will allow us to solve the general problem using a search algorithm. We assume that the values of $\mathbf{y}^*$ are nondecreasing and find the vector $\mathbf{a}$ which is nearest to $\mathbf{y}^*$ and satisfies the mean and variance constraint and has the additional properties that the first j_l's of the solution are are all equal and the last j_u's of the solution are also all equal and the difference between these two common values is r. Now a solution to this problem need not be a solution to the general problem since the range of the solution need not be r. In the next theorem we give the solution for this easier problem. The notation is the same as in the earlier Theorem 2.1.

Theorem 5.1 *Let $\mathbf{y}^*$ be vector of real numbers of length n and assume $y_i^* \leq y_j^*$ when $1 \leq i < j \leq n$ and where $n \geq 2$ is a fixed positive integer. Let m, $u > 0$ and $r > 0$ be fixed real numbers. Let j_l and j_u be fixed positive integers which satisfy $j_l + j_u < n$. Let $\mathbf{a}$ be a vector of real numbers of length n. For the problem of minimizing $\sum_{i=1}^n (a_i - y_i^*)^2$ over $\mathbf{a}$ subject to the constraints that $\bar{\mathbf{a}} = m$, $\mathrm{Var}(\mathbf{a}) = u$, $a_1 = \cdots = a_{j_l}$ and $a_1 + r = a_{n-(j_u-1)} = \cdots = a_n$ the solution is given by $\mathbf{a}^o$ where*

$$a_1^o = m - \frac{j_u}{j_l + j_u} r + \lambda^{1/2}(\bar{\mathbf{y}}^*(j_l, j_u) - \bar{\mathbf{y}}^*) \tag{5.1}$$

and for $i = j_l + 1, \ldots, n - j_u$

$$a_i^o = m + \lambda^{1/2}(y_i^* - \bar{\mathbf{y}}^*) \tag{5.2}$$

where

$$\bar{\mathbf{y}}^*(j_l, j_u) = (j_l + j_u)^{-1} \left(\sum_{i=1}^{j_l} y_i^* + \sum_{i=n-(j_u-1)}^{n} y_i^* \right) \tag{5.3}$$

and

$$\lambda = \frac{(n-1)u - (j_l + j_u)^{-1} j_l j_u r^2}{(j_l + j_u)(\bar{\mathbf{y}}^*(j_l, j_u) - \bar{\mathbf{y}}^*)^2 + \sum_{i=j_l+1}^{n-j_u} (y_i^* - \bar{\mathbf{y}}^*)^2} \tag{5.4}$$

Although not completely straightforward the proof of this theorem follows the proof of Theorem 2.1 and will be omitted. Note that this solution is

similar to the earlier solution if we keep in mind the form of the additional constraints involving j_l and j_u.

Now the solution vector $\mathbf{a}^o$ for this theorem need not satisfy the constraint that its range is r. Assuming the general problem has a solution one way to find it would be to compare all the solutions for all possible choices of j_l and j_u where the mean, variance and range constraints are all satisfied and find the one that is closest to $\mathbf{y}^*$. Since we believe that for problems with a continuous variable most statisticians would prefer a created sample with small values of j_l and j_u we will only consider created solutions given by the theorem where j_l and j_u range between 1 and some upper bound say K. If in this set of solutions none of the them satisfy the range constraint we just take the solution whose range is closest to r. In the following example we take K large enough so that even when we do not find the optimal solution the created sample sample selected exceeds the range constraint by only a small amount.

Consider again the population that was used to generate the results in Table 1. We repeated the simulation for the three cases except we just considered the two samples resp and adall. But now when we found the created sample adall we added the range constraint that its range should equal the range of the responders and used the search algorithm described above. The results are given in Table 3. Note that as the number of nonresponders increases the created solutions have more and more ties at the lower and upper ends of the solution. This is unsatisfactory particularly for a continuous variable. Moreover some additional values are moved quite close to the minimum value of the created sample and at the upper end others are moved quite close its maximum value. Although this does not seem to have much effect on the median of the created sample as an point estimator of the population median it does explain what at first glance are counter intuitive results of Tables 1 and 3. That is, even though in the simulations given in Table 3 the created samples have smaller overall ranges than those in Table 1 the average length of the confidence intervals for the median when $n_r = 25$ is considerable longer than the corresponding interval in Table 1.

From these results we see that as the number of nonresponders increases neither the solution given by Theorem 2.1 nor the solution of this section with the added range constraint is satisfactory. The first is two spread out while the second has too many ties near the ends. Each is a distortion in different ways of the information in the responders and would be found unacceptable by statisticians. These results also highlight the fact that as number of

constraints increase the created sample is likely to move further and further away from the responders. However if the number of nonresponders are small both will mimic the responders quite closely and for some purposes be almost as good as the responders.

put table 3 about here

6 Multivariate data

Suppose that associated with each unit in the population there is a collection of possible characteristics of interest. In this case y_i will be a vector. We assume that for each sampled unit some of its components could be missing. Given a sample where some of the y_i's have missing values the problem is to create a sample with no missing values and which when analyzed with standard frequentist methods yield correct inferential procedures. As for the earlier problem just imputing some values for the missing observations will not be good enough. In principle we would like a created sample which not only has the proper marginal properties for each characteristic but also preserves joint relationships between the characteristics as well. For example in addition to getting the first two moments "correct" marginally for each characteristic of interest we might want to have all the correlation coefficients in the created sample agree with the observed sample correlations. See the comments in Judkins (1996) on why this is an important and difficult problem. Unfortunately it is not clear how to do this. But we shall now argue that the essentially univariate approach described above can perform quite well. In particular if one can impute sensible values for the missing observations then adjusting all the values of each characteristic individually to get the first two moments right will also preserve the correlation structure.

To see why this is so let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ be a vector of real valued random variables with a known joint distribution for which $E(|\prod_{i=1}^m X_i|) < \infty$. Suppose when observing an outcome from this distribution the value $X_1 = x_1$ is always missing and we report $X_1^* = E(X_1|x_2, \ldots, x_m)$ in its place. Then for any nonempty subset s of $\{2, \ldots, m\}$ it is easy to see that

$$E(X_1^* \prod_{i \in s} X_i) = E(X_1 \prod_{i \in s} X_i) \tag{6.1}$$

Now suppose we have a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from this distribution. Moreover assume that for each $\mathbf{X}_i$ there is a positive probability that one

of its components could be missing at random. That is for each $\mathbf{X}_i$ we either observe the entire vector or we see all of its components but one. The missing component may vary from trial to trial. Finally suppose that when a component is missing we put in its place its conditional expectation given the rest of components. Let $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$ denote the random sample where the missing values have been replaced by their conditional expectations. Let s be a nonempty subset of $\{1, \ldots, m\}$, $W = \prod_{i \in s} X_i$ and $W^* = \prod_{i \in s} X_i^*$. It follows from the above remarks that for a sample of the $\mathbf{X}_i$'s both the sample mean of the W_i's and the sample mean of the W_i^*'s will be unbiased estimators of $E(W)$. Note however their variances will be different. Still this imputation scheme is preserving a good bit of the structure that is in the data. In particular the $\mathbf{X}_i^*$'s should yield a reasonable point estimator for the correlation coefficient for any pair of components of $\mathbf{X}$.

Now suppose we take the x_1^*'s from our sample and find the created sample closest to it which has the proper mean and variance so that marginal inferences about $E(X_1)$ based on it are the same as those which just uses the responders in the x_1^*'s. In the same way can marginally get created samples for every component of $\mathbf{X}^*$. Since for each component the new values in the created sample are just a linear transformation of the observed and imputed values of our completed sample the created sample will have the same correlation structure as the completed sample which contains the observed and imputed values. Even though the linear transformations used for the components will in general be different this process must preserved all the pairwise correlations. On the other hand if we just adjusted the imputed values in each component the correlation of the created sample can change by a fair amount even though we are just changing a few values. Now in practice one would never know the necessary conditional distributions to implement the above method. We will consider a bivariate example where the correlation coefficient is to be estimated. We will see that even with a naive imputation scheme adjusting all the values in the completed sample works much better than adjusting the imputed values and performs almost as well as using the sample points with no missing components.

Suppose $y_i = (y_{i_1}, y_{i_2})$ and consider a sample point i^* where $y_{i_2^*}$ is missing. We then divide the sample points with no missing values into two groups. The first is all those sample points whose value for the first characteristic of interest is less than or equal to $y_{i_1^*}$ while the second is those where this value is greater than $y_{i_1^*}$. In the first group we find the sample point with the largest value of y_{i_1} and in the second group we find the sample point with

the smallest value of y_{i_1} and then impute the average of their y_{i_2} values to $y_{i_2^*}$. If only one of these two groups is nonempty then we just impute the corresponding y_{i_2} value. Note this is quite naive but intuitively reasonable. However the properties of frequentist procedures will be altered by such imputations. To overcome this we will marginally construct created samples for each characteristic individually. One could either adjust every value in the sample or just the imputed values. To see how this would work in practice we considered the population used in the earlier simulations except that now we assumed that the x and y variables were both characteristics of interest.

We did two different simulations where each consisted of 500 random samples. For the first group the sample size was 50 and five units in the sample were selected at random and had the value of the first characteristic deleted and then from the remaining 45 we selected five more at random and deleted the second characteristic. In the second the sample size was 25 and the number deleted was two in each case. Then imputed values were assigned to the missing values using the procedure described just above. Then two created samples were formed. In one all the values were adjusted to satisfy a range constraint and to have the 'proper' mean and variance. In the other just the imputed values were adjusted subject to the three constraints. We denote these two methods as adall and adimp respectively. Then for the full sample, the sample of responders, i.e. the subset of the sample with no missing values, adall and adimp the sample correlation and its absolute deviation from the true population correlation were found. The results are given in Table 4. As was to be expected using the responders does just about as well as the full sample since the percentage of units in the sample with one characteristic missing was 20% and 16% respectively. The adall method seems to work almost as well as using just the responders while the method which just adjusts the imputed values performs significantly worse. Although some may find this quite surprising we believe it is explained in the argument given above.

put table 4 about here

7 Discussion

The Bayesian approach to the problem of nonresponse is quite natural from a theoretical point of view since in principle one can find a predictive distribution for the unobserved members of the population given the sample.

In particular one has a predictive distribution of the nonresponders given the responders in the sample. Practically it is more difficult however since in addition to selecting a prior distribution one must model the mechanism of nonresponse. Even if one assumes that the nonresponse is ignorable one still must simulate from this predictive distribution for the nonresponders to make inferences. Although someday this may not be considered a real difficulty for now it is still a formidable task for the typical user of public-use files. Multiple imputation which is a combination of Bayesian and frequentist ideas is a solution for this practical problem which is an improvement over single imputation methods.

The best frequentist answer to this problem is that found in Rao and Shao (1992), Rao (1996) and Fay (1996). In this approach the statistician must construct a single completed sample but then uses a jackknife type estimate of variance which treats the imputed values differently than the observed values to get a "correct" estimate of variance. This is an improvement over methods which just ignore the fact that some values have been imputed but has the added complication of treating the imputed values differently than the rest of the sample.

Binder (1996) notes that although deterministic imputation makes it possible to find appropriate estimates of variance when estimating the mean and the imputed values are identified such an approach may lead to biased results when estimating nonlinear quantities. He suggests for such problems stochastic imputation could be used. However the model used to generate the nondeterministic imputed values may also lead to biased results. In all the examples considered here we have used simple but hopefully fairly sensible methods of imputation. We have considered both deterministic and random imputation methods. In this approach the method of imputation is not so crucial since the created sample is selected so that inferences for the population mean will be correct. This results in a created sample which will always mimic certain selected properties of the sample of responders. On the other hand a very poor imputation procedure will result in distorted created samples which could work poorly for estimating quantities other than the mean.

The approach advocated here also has the advantage of dealing with just a single sample where all the values can be treated in the same manner. In all the simulations except those involving the ratio estimator in section 4 the created sample which was formed by adjusting all the values in the sample performed better than the created sample which was formed by adjusting

just the imputed values and even there the performances of the two methods were nearly identical. This is somewhat counter intuitive and I believe that it could cause many statisticians some unease. To see why this need not always be so recall that in many cases the sample is of interest only for what it tells us about the population it was drawn from. In our simulations the created samples only seemed to lose significant amounts of information about the tails of the population. This suggests for many purposes a created sample can be almost as good as the original. On the other hand there are situations where this would not be so. Suppose that in addition to making inferences about the population we are interested in the particular units making up the sample for some other purpose. For example we might want to pick out the best ones to be used in some future experiment. Then clearly a created sample should not be used. Since in this case the values attached to a particular unit carry important information about that unit. However when the sample is of interest only as a surrogate for the entire population then created samples should not be a cause for concern as long as they will yield proper inferences about the population.

Multiple imputation, the approach of Rao and Shao and the approach given here were all developed with the goal of finding a good frequentist confidence interval for the population mean when some of the observations in the sample are missing at random. One drawback of these approaches is that they may yield inappropriate inferences for other problems. We have presented several simulation studies that indicate that an appropriate created sample which adjusts both the observed and imputed values in the sample can yield sensible frequentist answer in a variety of problems where the characteristic of interest is a continuous variable. Moreover the created samples seem to preserve much of the information about the population contained in the responders as long as the number of nonresponders is not too large. When there are two many nonresponders however a created sample can be a gross distortion of the information in the sample even though it mimics a few selected statistics of the responders. On the other hand the simplicity of this approach along with the fact that it does not depend strongly on the method used to impute the missing values suggests that in some cases it can be a reasonable compromise solution for problems where a single completed data set is desired.

References

Basu, D. (1971). An essay on the logical foundations of survey, part one. In Godambe, V. P. and Sproutt, D. A., editors, *Foundations of Statistical Inference*. Hold, Rinehardt and Winston, Toronto.

Binder, D. A. (1996). Comment on the papers by Rubin, Fay and Rao. *Journal of the American Statistical Association*, 91:510–515.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91:490–498.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87:533–540.

Ghosh, M. and Meeden, G. (1997). *Bayesian methods for finite population sampling*. Chapman & Hall, London.

Judkins, D. R. (1996). Comment on the papers by Rubin, Fay and Rao. *Journal of the American Statistical Association*, 91:507–510.

Meeden, G. (1995). Median estimation using auxiliary information. *Survey Methodology*, 21:71–77.

Meeden, G. and Bryan, M. (1996). An approach to the problem of nonresponse in sample survey using the Polya posterior. In *Bayesian Analysis in Statistics and Econometrics Essays in Honor of Arnold Zellner*, pages 423–422. Wiley.

Meeden, G. and Ghosh, M. (1983). Chosing between esperiments: 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.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual intake distributions. *Journal of the American Statistical Association*, 91:1440–1449.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91:499–506.

Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79:811–822.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Table 1: The averages of the range, the median and the absolute error of the median for the different samples. Also the averages of the lower bound and the length of the Woodruff 95% interval estimate along with its frequency of coverage. In each of the three cases there were 500 random samples of size 50 and the number of responders were 45, 35 and 25.

Number of responders	Sample type	Ave range	Ave median	Ave error	Ave lowbd	Ave length	Freq of coverage
	full	34.0	43.9	1.03	41.6	4.92	0 .934
45	resp	33.5	44.2	1.18	41.6	5.42	0.950
	adall	36.6	44.1	1.12	41.5	5.29	0.948
	adimp	36.8	44.2	1.17	41.6	5.35	0.934
	resp	31.2	44.2	1.35	41.1	6.46	0.942
35	adall	41.2	44.1	1.26	44.1	6.10	0.958
	adimp	45.9	44.2	1.32	41.3	5.91	0.936
	resp	29.0	44.3	1.63	40.7	7.69	0.946
25	adall	49.2	43.9	1.37	40.4	7.23	0.940
	adimp	56.0	44.0	1.42	40.9	6.61	0.898

Table 2: The averages for 500 random samples of certain sample quantiles and the sample correlation for three different samples for a sample size of 50 with 40 responders.

Type	0%	25%	50%	75%	100%	Corr
resp	31.3	39.4	44.0	49.4	63.8	0.86
adall	29.6	39.5	44.2	49.3	65.1	0.83
adimp	30.0	39.3	44.0	49.5	65.0	0.83

Table 3: The averages of the median and its absolute error for estimating the population median for the sample of responders and a created sample when a range constraint is included. Also given are the averages of the lower bound and the length of the Woodruff 95% interval estimate along with its frequency of coverage for both samples. Note j_l and j_u are the number of lower and upper ties in the created sample. There were 500 random samples of size 50 and the number of responders were 45, 35 and 25.

Number of responders	Sample type	Ave median	Ave error	Ave lowbd	Ave length	Freq of coverage	Ave j_l	Ave j_u
45	resp	44.22	1.17	41.6	5.46	0.946		
	adall	44.15	1.12	41.5	5.53	0.948	1.92	1.42
35	resp	44.34	1.43	41.2	6.50	0.952		
	adall	44.01	1.30	40.8	6.89	0.972	4.02	3.46
25	resp	44.33	1.64	40.6	7.74	0.956		
	adall	43.12	1.80	37.89	11.1	0.980	11.7	8.71

Table 4: The averages of the sample correlation and its absolute error for 500 random samples of size $n = 50$ and $n = 25$ where each sample had respectively 5 and 2 missing values for each characteristic. The population correlation is 0.87.

	$n = 50$		$n = 25$	
Type	Ave value	Ave error	Ave value	Ave error
all	0.86	0.029	0.86	0.043
resp	0.86	0.030	0.86	0.045
adall	0.87	0.032	0.87	0.048
adimp	0.80	0.065	0.76	0.11