Recent Developments in Fractional Imputation

Wayne A. Fuller[1]

[1]Department of Statistics, Iowa State University, Ames, IA 50010

Abstract
Kalton and Kish (1984) suggested fractional imputation as an efficient hot deck imputation method. Fully efficient fractional imputation is the limit case in which the estimated conditional distribution is imputed for each missing item. Improved computational power has made fully efficient fractional imputation possible in many cases. Sampling of respondents is required in others. We discuss developments in replication variance estimation, in parametric imputation, approximations to fully efficient fractional imputation, and extensions to multivariate imputation.

Key Words: Missing data, hot deck, multivariate imputation, random imputation.

Research on fractional imputation dates from the work of Kalton and Kish in the 1980's. The monograph of Kalton (1983) contains a discussion of several aspects of adjusting for missing data, "Compensation" as Kalton called it. He listed three important attributes of a good procedure. First, the estimates should have good precision and he has considerable discussion on the use of sampling techniques in selecting donors. The second of his criteria concerns standard error estimation. He pointed out that recording donors is important for standard error estimation. Kalton called the third criterion, "general applicability". (Kalton 1983, p. 32). Any compensation procedure should be suitable for different parameters because the data will be used for many analyses. That is, the imputed data set should provide a consistent estimator of the distribution function of any variable.

Kalton's discussion of general applicability illustrates one of the ways that people approach the problem of missing data. Kalton had responsibility for producing a data set whose uses were imperfectly known. In such a situation, one is certain that people will compute estimates for parameters that were not considered at the time of data set construction. Those with responsibility for a general purpose data set look for procedures that rely on as few assumptions as possible and that have the widest possible applicability. Efficiency is secondary to general applicability.

Subject matter specialists or those who work with them often begin by constructing a model for some of the data and looking for efficient estimation schemes. The outlook of such individuals differs from that expressed by Kalton because of the different primary responsibilities. People can have both types of responsibilities, but one's primary task dominates one's view of imputation and compensation.

A very simple missing data illustration is given in Table 1. There are five observations, and two variables, x and y. Assume the sample is a simple random sample. If we are estimating the mean, each element will have a weight of 0.20. The y is missing on observation 5 and the variable x can be used to define imputed values. Notice that x has two values; 1 and 2. The missing value has an x of 2. Hence, a reasonable procedure would be to choose at random one of the 3 (x,y) pairs where x is equal to 2 and place the chosen y value in the missing spot. If we chose observation 2 then the missing y value would be replaced with 1.

Table 1. Missing Data Illustration

Observation	w_i	x_i	y_i
1	0.2	1	1
2	0.2	2	1
3	0.2	2	1
4	0.2	2	2
5	0.2	2	Miss

To increase efficiency of the imputation estimator of the mean, Kalton and Kish (1981,1984) suggested that one assign *M* donors to each of the non-respondents and assign a weight of 1/*M* times the original weight to each of the donated values. They discuss methods of selecting donors to give good efficiency. For example, if we have n_m missing, we can construct an efficient estimator of the mean by selecting Mn_m donors and assigning *M* to each respondent. Then the grand mean estimate would be close to the mean of the respondents because we would use each potential donor nearly an equal number of times.

In the example of Table 1, assume that we set $M = 2$ and that we chose units 2 and 4 as our donors. Then each donated value would have a weight of one half of the original weight. The new data set has one more line of data. The total weight for observation 5 stays the same but we have a more efficient estimator of the mean of *y* because, instead of a single donor, we have two.

We can carry the idea of multiple donors a bit further and use all available donors for each recipient. We call the resulting estimator "fully efficient." In our example we use all three donors. Our estimator for the missing observation is the observed cumulative distribution function. Each of the three possible donors get one third of the total weight. The total weight for observation 5 remains 0.20, and the imputed "observation" contains all that we know about observations with $x = 2$. Because there are only two possible values for *y* we can simplify the data set. Only two rows are required. One row for observations of the type $(x,y) = (2,1)$ and one row for observations $(x,y) = (2,2)$. See Table 2.

Table 2. Missing Discrete Data Fully Efficient Imputation

Observation	w_i	w_{ij}^*	w_i^*	x	y
1	0.2	1.00	0.200	1	1
2	0.2	1.00	0.200	2	1
3	0.2	1.00	0.200	2	1
4	0.2	1.00	0.200	2	2
5*	0.2	0.67	0.133	2	1
5*	0.2	0.33	0.067	2	2

If we carry out imputation in this manner, we have a data set "generally applicable". For example, estimates for a domain are fully efficient in the sense that the missing data are suitable for estimation of the cumulative distribution function of *y* for the domain. We should say that the estimation is fully efficient in a class of imputation estimators.

To this point we have not explicitly specified the model basis for our imputation. A classical model uses the population divided into G cells. In our example there were two cells; one with $x = 1$ and one with $x = 2$. The y values can be decomposed into a mean for the cell plus an error, and a model is

$$y_{gi} = \mu_g + e_{gi} \doteq \mu_{gi} + e_{gi}$$
$$e_{gi} \sim ii(0, \sigma_g^2),$$

where we add the assumption that the errors are *iid*. The model assumes that the representation is appropriate for both respondents and nonrespondents. Given the assumptions, we can use the respondents to estimate the cell mean.

The imputation procedure we have been using is called hot deck imputation. The missing value is replaced with a value that exists in the data set. Hot deck imputation has the "general applicability" property that the imputed value is known to exist in the population.

The estimated mean using fractional imputation is a weighted average of the respondent values. The weight for respondent *i* is the original weight plus the sum of the fractional weights for respondent *i* donating to recipient *j*. The estimator for a simple random sample is

$$\hat{\mu}_I = n^{-1} \sum_{j \in A} \sum_{i \in A_R} d_{ij} w_{ij}^* y_i$$
$$= n^{-1} \sum_{i \in A} \mu_i + \sum_{i \in A_R} \alpha_i e_i$$

where

$$\alpha_i = n^{-1} \sum_{j \in A} d_{ij} w_{ij}^*$$

w_{ij}^* = the fractional weight, d_{ij} is an indicator with d_{ij} = 1 if respondent *i* donates to recipient *j* and is zero otherwise, *A* is the index set for the sample, A_R is the index set for the respondents, and A_m is the index set for the recipients.

We can combine the weights for a respondent so that the grand mean is the mean of the cell means plus a weighted average of the *e*'s. It follows that the variance of a fractionally weighted estimator is the variance of the mean of the individual cell means plus the expected value of the variance of the weighted average of the *e*'s. That is,

$$V\{\hat{\mu}\} = V\{\overline{\mu}\} + E\left\{ \sum_{i \in A_R} \alpha_i^2 \sigma_i^2 \right\}$$

where

$$\hat{\mu} = n^{-1} \sum_{i \in A} \mu_i + \sum_{i \in A_R} \alpha_i e_i$$

and

$$= \overline{\mu} + \sum_{i \in A_R} \alpha_i e_i .$$

The terms in $V\{\hat{\mu}\}$ can be estimated. If we're constructing a general use data set then replication procedures, such as the jackknife, are preferred variance estimation

procedures. For our imputation procedure, a natural way to construct jackknife replicates would be; delete a respondent and delete the corresponding imputed values. Then, increase the weights for remaining units and increase the fractional weights for the remaining donors so the sum of the fractions is one. This direct estimation procedure is biased because of a "degrees of freedom" problem. The jackknife for a simple random sample of size *n* requires a correction factor of $(n-1)/n$. If we delete one of the *M* respondents, something like an $M(M-1)/n$ adjustment would be appropriate for that individual recipient. Given a more complex sample or a more complex donor procedure, the adjustment is more complicated, but the basic requirement for adjustment remains. The bias in the direct procedure is small for large *n* and large *M*. Using the cell mean model, it is possible, but difficult, to construct jackknife weights for unbiased variance estimation. See Kim and Fuller (2004).

In large surveys, one may choose to sample donors rather than use the fully efficient procedure. As mentioned earlier, Kalton and Kish (1984) addressed the problem of donor selection. For example, by defining *M* strata of donors and selecting one from each stratum, one can create an efficient sample for each respondent. In the same way that was discussed by Kalton and Kish (1984), we can assure marginal efficiency by balancing the sample of respondents across recipients. One way to do this is use a rejective procedure Fuller (2009) in the selection of donors or to use the regression estimator.

Kim (2011) showed that fractional imputation provides an efficient way to estimate parametric models for data sets with missing items. Assume we have a parametric model for the data and a model for the response. One way to estimate the model is to start with a set of donors for each missing value with fractional weights that sum to one. Using those donors to define a complete data set, estimate the parameters of the model. Using the estimated parameters, update the fractional weights on the donated values, and iterate. Note that the donated values need not change at each step, only the fractional weights. To illustrate those ideas, we use the data set of Table 3 with 10 observations and three variables. The y_1 is missing for observation 3, y_3 is missing for 6, and both y_2 and y_3 are missing for observation 7. Observation 7 has y_1 value of 2 with (y_2, y_3) missing. There are four possible donors for observation 7 with three unique vectors. The unique vectors are (2,1,2), (2,1,1) and (2,2,1).

Table 3. Multivariable Data Set with Missing Items

Unit	y_1	y_2	y_3
1	1	1	1
2	2	1	2
3	M	2	1
4	2	1	1
5	2	2	1
6	1	1	M
7	2	M	M
8	2	2	1
9	1	1	2
10	1	2	1

Table 4. Estimated Probabilities for Observed Vectors

(y_1, y_2, y_3)	Fraction of Respondents	Probability
(1, 1, 1)	0.143	0.150
(1, 1, 2)	0.143	0.150
(1, 2, 1)	0.143	0.128
(2, 1, 1)	0.143	0.121
(2, 1, 2)	0.143	0.121
(2, 2, 1)	0.286	0.330

There are six unique values for the vector (y_1, y_2, y_3) in the data set, listed in Table 4. The fraction of the complete respondents in each category is given in column two. There is one observation for each of six vectors except the category (2,2,1), where two of the original seven respondents have the value. The third column is the estimated probability for each type of respondent calculated by the iterative procedure of Kim (2011). The vector (2,2,1) has an estimated probability greater than one seventh.

Using the estimated probabilities, the fractional weights are 0.21, 0.21, and 0.58 for the three possible imputed values for unit seven. (Table 5) If we had used the weights from the respondents, the fractions would be 0.25, 0.25, and 0.50.

Table 5. Unit Seven Imputed

Unit	Frac. Wt	y_1	y_2	y_3
7	0.21	2	1	1
7	0.21	2	1	2
7	0.58	2	2	1

Table 6 contains the final data set with imputed values and the fractions. The imputed values are starred in the table.

Table 6. Imputed Data Set

Unit	Frac. Wt	y_1	y_2	y_3
1	1	1	1	1
2	1	2	1	2
3*	0.28	1	2	1
3*	0.72	2	2	1
4	1	2	1	1
5	1	2	2	1
6*	0.50	1	1	1
6*	0.50	1	1	2
7*	0.21	2	1	1
7*	0.21	2	1	2
7*	0.58	2	2	1
8	1	2	2	1
9	1	1	1	2
10	1	1	2	1

The outlined estimation-imputation procedure is available in SAS®. See SAS Institute (2015). The computation is exactly as described. The jackknife weights are available, where those jackknife weights will yield somewhat biased estimates of variance.

References

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure, *Biometrika*, 96, 1-12.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Survey Research Center, University of Michigan, Ann Arbor, Michigan.

Kalton, G. and L. Kish (1981). Two efficient random imputation procedures. *Proceedings of the Section on Survey Research Methods, American Statistical Association.* 146-151.

Kalton, G. and L. Kish (1984). Some efficient random imputation methods. *Comm. Statist. Theory Methods. 13*(16), 1919-1939.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika 98*, 119-132.

Kim, J. K. and W. A. Fuller (2004). Fractional hot deck imputation. *Biometrika 91*(3), 559-578.

Kim, J. K., W. A. Fuller, and W. R. Bell. (2011). Variance estimation for nearest neighbor imputation for U.S. census long form data. *Ann. Appl. Stat. 5*(2A), 824-842.

Kim, J. K., A. Navarro, and W. A. Fuller (2006). Replicate variance estimation after multi-phase stratified sampling. *J. Amer. Statist. Assoc. 101*, 312-320.

SAS Institute Inc. (2015). SAS/STAT 14.1 User's Guide – the SURVEYIMPUTE procedure, Cary, NC: SAS Institute Inc.

Yang, S. and J. K. Kim (2016). Fractional imputation in survey sampling: A comparative review. *Statistical Science.* To appear.