Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 4, pp. 993~1002

Fully Efficient Fractional Imputation for Incomplete Contingency Tables

Shin-Soo Kang[1)]

Abstract

Imputation procedures such as fully efficient fractional imputation(FEFI) or multiple imputation(MI) can be used to construct complete contingency tables from samples with partially classified responses. Variances of FEFI estimators of population proportions are derived. Simulation results, when data are missing completely at random, reveal that FEFI provides more efficient estimates of population than either multiple imputation(MI) based on data augmentation or complete case analysis, but neither FEFI nor MI provides an improvement over complete-case(CC) analysis with respect to accuracy of estimation of some parameters for association between two variables like $\theta_{i+}\ \theta_{+j} - \theta_{ij}$ and log odds-ratio.

Keywords : Complete Case Analysis, Multiple Imputation

1. Introduction

In the analysis of contingency tables, it may happen that some observations are not fully cross-classified. This issue has been studied for a long time. One simple approach, known as complete-case(CC) analysis, discards the missing data by restricting analysis to only fully classified counts in an incomplete contingency table.

An alternative approach involves constructing a complete table, in which all cases are completed classified, by imputing information for the missing row or column classification. Multiple imputation, proposed by Rubin (1978), provides a way to take advantage of commonly used tests of independence for completely classified tables.

---

1) Professor, Department of Information Statistics, Kwandong University, Kangnung, 210-701, Korea
E-mail: sskang@kd.ac.kr

Fractional imputation was discussed by Kalton and Kish (1984) and Fay (1996). It is a hot deck imputation procedure that uses more than one responding unit as a donor for a missing unit. The fully efficient fractional imputation(FEFI) procedure described by Kim and Fuller (2004) uses every responding unit within a designated imputation group as a donor for a unit with missing information. This provides a single completed table and has some practical advantages over MI in that a single completed table can be published for public use.

Little and Rubin (2002) discuss three general mechanisms for missing data: missing completely at random(MCAR), missing at random(MAR), and not missing at random(NMAR).

Let X_1 and X_2 denote categorical variables for a two-way incomplete contingency table. If the missing probability of X_i does not depend on either the value of the other variable or the value of X_i, then it is MCAR. If the missing probability of X_i depends on the value of the other variable but not on the value of X_i, then it is MAR. If the missing probability of X_i depends on its value, then it is NMAR.

FEFI is easily implemented in two-way incomplete contingency tables when we assume that the mechanisms that lead to missing dat are 'MCAR'. FEFI imputes a number of values for the missing information on each partially classified observation along with a set of weights.

FEFI, and MI are reviewed in section 3. Small sample relative efficiency and bias are examined through Monte Carlo simulation in section 4. Methods for obtaining covariance matrices for FEFI estimates of population proportions are discussed in section 5.

2. Notation

Consider an $I \times J$ contingency table where the row factor X_1 has I categories and the column factor X_2 has J categories. Assume simple random sampling with replacement. In a complete table, where the row and column categories are observed for every case in the sample, the counts have a multinomial distribution with sample size N and probability vector θ. Let n_{ij} denote the count for the (i, j) cell, and let θ_{ij}, an element of θ, denote the population proportion for the (i, j) cell.

When information on either the row or column classification is missing, we can construct a table of counts for the completely classified cases where x_{ij} denotes the number of cases observed in the (i, j) cell. We can also construct one-way

tables of counts for partially classified cases. Let x_{im} denote the number of cases in the i^{th} row category, $i=1,2,\cdots,I$, where the column category is unknown, and let x_{mj} denote the number of cases in the j^{th} column category, $j=1,2,\cdots,J$, where the row category is unknown. Then, x_{im} and x_{mj} are marginally observed counts on a single variable. Let x_{mm} denote the number of cases where both the row and column categories are missing. The total sample size is

$$\begin{aligned} N &= \sum_{ij} x_{ij} + \sum_{i} x_{im} + \sum_{j} x_{mj} + x_{mm} \\ &= n_{cc} + x_{+m} + x_{m+} + x_{mm}. \end{aligned}$$

3. Estimates of the population proportions

3.1 FEFI estimates under MCAR

Fully efficient fractional imputation(FEFI) is a kind of hot deck imputation which uses every responding unit as a donor for a missing unit within any particular imputation group. For a two-way contingency table obtained from a sample with no auxiliary variables, the imputation group for a unit with observed value of the row factor X_1 but missing X_2, the value of the column factor, is the set of complete cases with the same value of X_1. Similarly, the imputation group for a unit with missing value of X_1 is the set of complete cases with the same value of X_2. This simplifies the implementation of FEFI.

For a unit with only X_2 missing, imputation fractions for the J possible values of X_2 are obtained from the conditional frequencies of X_2 in the cross-classified table of complete cases given the observed value of X_1. The analogous procedure is used for any unit with only X_1 missing. For a unit with missing information for both X_1 and X_2, we impute $I\times J$ possible values with imputation fractions corresponding to the joint frequencies of X_1 and X_2 incorporating all partial information.

For example, consider a 2×2 incomplete contingency table where X_1 and X_2 assume values of 0 or 1 from a simple random sample of size $N=88$ and assume a completely missing at random mechanism. Table 1 shows observed counts for the 9 possible response patterns, using '?' to indicate a missing value for the corresponding variable.

Table 1: Response patterns for a 2×2 incomplete contingency table

X_1	1	1	0	0	?	?	1	0	?
X_2	1	0	1	0	1	0	?	?	?
Counts	5	10	15	20	8	9	6	7	8

Imputed information for the 8 observations with pattern (?,1) is btained from the relative frequencies of the (1,1) and (0,1) responses as shown in Table 2. The imputation fraction for (1,1), given the (?,1) response pattern is $x_{11}/x_{+1}=5/(5+15)=0.25$, which corresponds to allocating $8\times0.25=2$ counts to (1,1) from (?,1), and allocating 6 counts to (0,1) from (?,1). This procedure is repeated for the (?,0), (1,?), and (0,?) patterns yielding the allocated counts shown in Table 2.

Table 2: FEFI for partially classified cases in Table 1

X_1	?		?		1		0	
X_2	1		0		?		?	
Counts	8		9		6		7	
Allocation	(1,1)	(0,1)	(1,0)	(0,0)	(1,1)	(1,0)	(0,1)	(0,0)
FEFI	2	6	3	6	2	4	3	4

The updated complete table shown in Table 3, incorporates all partial information; it does not include the units with missing information on both variables. The 8 observed counts in (?,?) are allocated to (1,1), (1,0), (0,1) or (0,0) with respect to the relative frequencies 9/80, 17/80, 24/80, 30/80 obtained from Table 3. The resulting FEFI allocations from (?,?) are shown in Table 4. The completed table provided by FEFI is shown in Table 5.

Table 3: Updated table using complete and partial complete cases

X_1	1	1	0	0	?
X_2	1	0	1	0	?
Counts	9	17	24	30	8

Table 4: FEFI for cases with no information

X_1	?			
X_2	?			
Counts	8			
Allocation	(1,1)	(0,1)	(1,0)	(0,0)
FEFI	0.9	2.4	1.7	3

Table 5: FEFI completed table based on N

X_1	1	1	0	0
X_2	1	0	1	0
Counts	9.9	18.7	26.4	33

The counts in the completed table obtained by fully efficient fractional imputation under simple random sampling are given by the following formula:

$$\begin{aligned}\hat{n}_{ij}^{*} &= x_{ij}+x_{ij}\left(\frac{x_{im}}{x_{i+}}+\frac{x_{mj}}{x_{+j}}\right)+\frac{x_{ij}x_{mm}}{N-x_{mm}}\left(1+\frac{x_{im}}{x_{i+}}+\frac{x_{mj}}{x_{+j}}\right) \\ &= x_{ij}\left(1+\frac{x_{im}}{x_{i+}}+\frac{x_{mj}}{x_{+j}}\right)\left(1+\frac{x_{mm}}{N-x_{mm}}\right),\end{aligned} \tag{1}$$

where x_{ij} is the observed count for fully observed cases prior to imputation, $x_{i+}=\sum_{j=1}^{J}x_{ij}$ and $x_{+j}=\sum_{i=1}^{I}x_{ij}$. The total sample size is $N=\sum_{ij}\hat{n}_{ij}^{*}$. Discarding the x_{mm} cases for which both variables are missing does not affect the relative allocation in the completed table. The relative allocation in Table 5 is the same as the relative allocation in the first four columns of Table 3. Those cases do not contain any information about the joint distribution of X_1 and X_2. Therefore (1) can be simplified as

$$\hat{n}_{ij}=x_{ij}\left(1+\frac{x_{im}}{x_{i+}}+\frac{x_{mj}}{x_{+j}}\right),$$

with N changed to $n=N-x_{mm}$. The FEFI estimates of the population proportions are

$$\hat{\theta}_{FEFI} = \frac{1}{n}(\hat{n}_{11}, \hat{n}_{12}, \cdots, \hat{n}_{1J}, \hat{n}_{21}, \hat{n}_{22}, \cdots, \hat{n}_{2J}, \cdots, \hat{n}_{IJ})'$$
$$= \frac{1}{N}(\hat{n}_{11}^{*}, \hat{n}_{12}^{*}, \cdots, \hat{n}_{1J}^{*}, \hat{n}_{21}^{*}, \hat{n}_{22}^{*}, \cdots, \hat{n}_{2J}^{*}, \cdots, \hat{n}_{IJ}^{*})'.$$

3.2 MI estimates

The multiple imputation(MI) procedure proposed by Rubin (1978) offers another possibility for estimation of cell probabilities for an incomplete contingency table. Suppose there are D imputed data sets. Each imputed data set is analyzed using the standard complete-data method. Let $\hat{\theta}_d$, $d=1,\cdots,D$ be the standard complete-data estimates for the vector of cell probabilities from the D imputed data sets. Then, MI estimates of the cell probabilities are given by

$$\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d.$$

4. Simulation Results

All of the 2×2 incomplete contingency tables for this study were generated with equal cell probabilities and data missing completely at random. Four combinations of sample size and level of missing data were considered and 1000 tables were generated for each combination. X_1 and X_2 were independently generated as Bernoulli(0.5) random variables. There are two levels 200 and 400 for the total sample size N. MX_i is a missing indicator variable independent of X_i. If $MX_i=1$, the corresponding variable X_i is missing. The four combinations of factors are summarized in Table 6. The percentages of cases with missing information on at least one variable are expected to be 19%, 36%, 51%, and 91% for combination 1, 2, 3, and 4, respectively.

Table 6: Combination of factors

Combination	N	$MX_i \sim Ber(p)$ p
1	200	0.1
2	200	0.2
3	200	0.3
4	400	0.7

θ_{11} and $\theta_{1+}\theta_{+1}-\theta_{11}$ were estimated by FEFI,MI, and complete-case analysis(CC). CC is a standard complete-data analysis discarding all missing cases. For multiple imputation, data augmentation with Jeffreys noninformative prior(Box and Tiao, 1992) was used to generate imputed data and construct 5 completed tables through S-PLUS 6.1(2001) functions for missing values. Mean and standard errors(S.E.) of 1000 values of the point estimates are shown in Table 7 and Table 8.

Table 7 shows means and standard errors of 1000 values for the estimates of θ_{11}. The true value of θ_{11} is 0.25. All four methods provide essentially unbiased estimates for the cell probabilities. The standard errors of the estimates differ across methods. Complete-case analysis provides the estimate of θ_{11} with the largest variance. For all methods, variation increases as the proportion of missing values increases. FEFI tends to provide smaller standard errors of cell proportion than MI in most cases, but the standard error of the MI estimator was smaller than FEFI for combination 4 which had largest proportion of missing values.

Table 7: Estimation of θ_{11}

Combination	FEFI		MI		CC	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
1	0.2488	0.0340	0.2488	0.0340	0.2489	0.0352
2	0.2489	0.0344	0.2487	0.0358	0.2491	0.0367
3	0.2483	0.0399	0.2485	0.0409	0.2490	0.0444
4	0.2513	0.0566	0.2521	0.0555	0.2518	0.0722

Table 8 shows means and standard errors of 1000 simulated values for the estimates of $\theta_{1+}\theta_{+1}-\theta_{11}$, a measure of association between the two variables, when the true value of $\theta_{1+}\theta_{+1}-\theta_{11}$ is 0. The averages of the estimates are similar for all methods, but the complete-case exhibits smaller standard errors than FEFI or MI.

Table 8: Estimation of $\theta_{1+}\theta_{+1}-\theta_{11}$

	FEFI		MI		CC	
Combi.	Mean	S.E.	Mean	S.E.	Mean	S.E.
1	-0.000011	0.02002	0.000033	0.02042	-0.000010	0.01999
2	-0.000082	0.02242	0.000154	0.02311	-0.000079	0.02235
3	0.000815	0.02495	0.000627	0.02603	0.000792	0.02479
4	-0.000493	0.04277	-0.001406	0.04479	-0.000650	0.04117

The imputation methods, FEFI and MI can not provide more information on association between two variables than what is present in the observed data. Although the imputation methods improve the estimation of individual cell probabilities relative to complete-case analysis, the covariance matrix for estimated cell probabilities is also affected by imputation.

5. Variance of FEFI Estimates

An estimate of the large sample variance-covariance matrix of the FEFI estimates of cell probabilities is derived using the delta method.

Let $C_0 = (x_{11}, \cdots, x_{1J}, x_{21}, \cdots, x_{2J}, \cdots x_{IJ}, x_{m1}, \cdots, x_{mJ}, x_{1m}, \cdots x_{Im})'$. Conditional on the value of x_{mm}, C_0 has a multinomial distribution with sample size $n = N - x_{mm}$ and probabilities

$$\pi = (\pi_{11}, \cdots, \pi_{1J}, \pi_{21}, \cdots, \pi_{2J}, \cdots \pi_{IJ}, \pi_{m1}, \cdots, \pi_{mJ}, \pi_{1m}, \cdots \pi_{Im})'.$$

The variance-covariance matrix of C_0 is $Var(C_0) = n(\Delta_\pi - \pi\pi')$, where Δ_π is a diagonal matrix with the elements of π on the main diagonal.
Each element of $\hat{\theta}_{FEFI}$ is a function of the elements of C_0. By the delta method, the variance of $\hat{\theta}_{FEFI}$ is derived as

$$Var(\hat{\theta}_{FEFI}) = \frac{1}{n} D(\Delta_\pi - \pi\pi')D' \equiv \Sigma_F,$$

where

$$D_{p\times q} = \begin{pmatrix} \frac{\partial \hat{n}_{11}}{\partial x_{11}} & \cdots & \frac{\partial \hat{n}_{11}}{\partial x_{IJ}} & \frac{\partial \hat{n}_{11}}{\partial x_{m1}} & \cdots & \frac{\partial \hat{n}_{11}}{\partial x_{mJ}} & \frac{\partial \hat{n}_{11}}{\partial x_{1m}} & \cdots & \frac{\partial \hat{n}_{11}}{\partial x_{Im}} \\ \frac{\partial \hat{n}_{12}}{\partial x_{11}} & \cdots & \frac{\partial \hat{n}_{12}}{\partial x_{IJ}} & \frac{\partial \hat{n}_{12}}{\partial x_{m1}} & \cdots & \frac{\partial \hat{n}_{12}}{\partial x_{mJ}} & \frac{\partial \hat{n}_{12}}{\partial x_{1m}} & \cdots & \frac{\partial \hat{n}_{12}}{\partial x_{Im}} \\ \vdots & \vdots & & & & & & & \vdots \\ \frac{\partial \hat{n}_{IJ}}{\partial x_{11}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial \hat{n}_{IJ}}{\partial x_{mJ}} & \frac{\partial \hat{n}_{IJ}}{\partial x_{1m}} & \cdots & \frac{\partial \hat{n}_{IJ}}{\partial x_{Im}} \end{pmatrix},$$

and $p = I \times J$, $q = I \times J + I + J$, with

$$\frac{\partial \hat{n}_{ij}}{\partial x_{cd}} = \begin{cases} 1+\left(\frac{x_{im}}{x_{i+}}+\frac{x_{mj}}{x_{+j}}\right)-x_{ij}\left(\frac{x_{im}}{x_{i+}^2}+\frac{x_{mj}}{x_{+j}^2}\right), & c=i \text{ and } d=j \\ -\left(x_{ij}\frac{x_{im}}{x_{i+}^2}\right), & c=i,\ d\neq j,\ \text{and } d\neq m \\ -\left(x_{ij}\frac{x_{mj}}{x_{+j}^2}\right), & c\neq i,\ d=j,\ \text{and } c\neq m \\ \frac{x_{ij}}{x_{+j}}, & c=m \text{ and } d=j \\ \frac{x_{ij}}{x_{i+}}, & c=i \text{ and } d=m \\ 0, & \text{otherwise}. \end{cases}$$

By the Central Limit Theorem and the Delta method, the FEFI estimators have an approximate multivariate normal distribution with expectation θ and variance Σ_F.

6. Discussion

Imputation using FEFI or MI provides more efficient estimates of cell probabilities than complete-case(CC) analysis. When data are missing completely at random and other covariates are not available, neither FEFI nor MI provides an improvement over complete-case(CC) analysis with respect to accuracy of estimation of some parameters for association between two variables like $\theta_{i+}\ \theta_{+j} - \theta_{ij}$ and log odds-ratio.

When data are missing completely at random, FEFI is easier to implement than MI. Explicit formulas for estimates of cell probabilities are given in (1). If the missing mechanism does not satisfy missing completely at random(MCAR) criterion, complete-case(CC) analysis can produce biased estimates of joint probabilities and distorted p-value for tests of independence. The FEFI method described in this article yields consistent estimators of cell probabilities if the missing mechanism is MCAR, but it is not necessarily consistent if data are simply missing at random(MAR). The allocation of the partially classified counts must be modified for fully efficient fractional imputation(FEFI) to provide consistent estimates for the MAR situation. MI provides consistent results under either the MAR situation.

Another approach to estimation of joint cell probabilities that can be applied when the missing mechanism is either MCAR or MAR is maximum likelihood estimation using both the complete and partially classified cases. Little(1982) developed a simple EM algorithm for two way contingency tables. This approach does as well as FEFI in estimation of the joint probabilities, but variance

estimation is more complicated.

References

1. Box, G. E. P., and Tio, G. C. (1992). *Bayesian Inference in Statistical Analysis*, (Wiley Classics Library Edition). J. Wiley and Sons, New York.
2. Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data, *Journal of the American Statistical Association*, 91, 490-498.
3. Kalton, G., and Kish, L. (1981). Two Efficient Random Imputation Procedures, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 146-151.
4. Kim, J. K., and Fuller, W. (2004). Fractional Hot deck imputation, *Biometrika*, 91, 559-578.
5. Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
6. Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data*, J. Wiley and Sons, New York.
7. Rubin, D. B. (1978). Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse, *Proceedings of the Survey Research Methods Section*, *American Statistical Association,* 1978, 20-34.
8. S-Plus 6.1 Manual (2001). *Analyzing Data with Missing Values in S-Plus,* Insightful Corporation. Seattle, Washington.

[received date : Sep. 2004, accepted date : Nov. 2004]