A Missing Technique for Estimating Missing Values

Silvia Irin Sharna, Mian Arif Shams Adnan and Rahmatullah Imon
Ball State University, 2000 W University Avenue, Muncie, IN 47306

Abstract

A missing value resembles not necessarily only the unknown data of an unknown probability of distribution but also their unknown characteristics. In this situation, it is better to construct a basket of characteristics based on assumed missing values. So, our immediate objective will be prudently picking the core characteristics for estimating the missing values and we refer this technique a "Missing" Technique.

Key Words: Average Log Likelihood Function, Combination, Dummy Missing Value, Likelihood Rate, Simple Random Sample.

1. Introduction

Missing data pattern describes which values are observed in the data matrix and which values are missing, and missing data mechanism addresses the relationship between missing value and the available values in the data matrix. Missing value estimation is a common problem in several statistical studies. The problem synchronized a lot when the sample size is very small and sensitive. Missing data mechanisms demonstrate the dependencies among the missing data and the available data. Rubin (1976) developed a device of treating the missing data indicators as random variables along with a distribution. The literature on analysis of partially missing data is inaugurated by Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird, and Rubin (1977), Litte and Rubin (1983), Little and Schenker (1994), and Little (1997). Methods proposed by the aforesaid authors can be grouped into the following categories.

The categories include Procedures Based on Completely Record Units, Weighting Procedures, Imputation-Based Procedures and Model-Based Procedures.

When some variables are not recorded for some of the units, the method analyzes only the units of the competed data (e.g., Nie et al., 1975). It can lead serious biases, however, and it is not usually very efficient, especially when drawing inferences for subpopulations. The weighting procedure demonstrates the randomization inferences from sample survey data without nonresponse commonly weight sampled units by their design weights, which are inversely proportional to their probabilities of selection. The missing values are filled in and the resultant completed data are analyzed by standard methods. Commonly used procedures for imputation include hot deck imputation, where recorded units in the sample are used to substitute values; mean imputation, where means from sets of recorded values are substituted; and regression imputation, where the missing variables for a unit are estimated by predicted values for the regression on the known variables for that unit. A broad class of procedures is generated by defining a model for the observed data and basing inferences on the likelihood or posterior distribution under that model, with parameters estimated by procedures such as maximum likelihood.

Broadly there are two ways for estimating missing values. These are Missing Value Estimation in Experiment and Missing Value Estimation by Likelihood Based Method. Imputation Method, Weighted Methods by Complete Case and Available Case Analysis are from class one. And Inference based Likelihood method, Factored Likelihood Method, EM Algorithm, Large Sample Inference based Maximum Likelihood Method, Bayesian Iterative Simulation Method, Robust Method, Partially Classified Contingency Table Method (ML Estimation, Bayes Estimation, Log-linear Model, Logistic Regression Method) etc. are from class two.

Missing data estimation in experiments includes Least Square analysis with missing data using Yates Method, Allan and Wishart's (1930) method for the Least square estimate of one missing value in the Randomized Block Design and of one missing value in a Latin Square Design, Wilkinson's (1958) method by providing formulas for many designs and many patterns of missing value, Hartley's (1958) non-iterative Method for estimating one missing value to be used iteratively for more than one time. The method for one missing value involves substituting three different trial values for the missing value, with the residual sum of squares calculated for each trial value. Since the residual sum of squares is quadratic in the missing value, the minimizing value of the one missing value can be found. The method is not as attractive as an alternative method. Healy and Westmacott (1956) described a popular iterative method with five steps. In step one, the trial values are substituted for all missing values, at step two the complete data analysis is performed, predicted values are obtained for the missing values at step three, in step four these predicted values are substituted for the missing values, a new complete data analysis is performed and so on until the missing values do not change appreciably or equivalently until the residual sum of squares essentially stops decreasing. In some cases, convergence can be slow and special acceleration techniques have been suggested by Pearce (1965). Although these can improve the rate of convergence in some examples, they can also destroy the monotone decrease of the residual sum of squares in other examples. A general non-iterative method due to Bartlett (1937) is to fill in guesses for the missing values, and then perform an analysis of covariance (ANCOVA) with a missing value covariate for each missing value. The *i*-th missing value covariate is defined to be indicator for the *i*-th missing value, that is, zero everywhere except for the *i*-th missing value where it equals one. The coefficient of *i*-th missing value covariate, when subtracted from the initial guess of the *i*-th missing value, yields the least square estimate of the *i*-th missing value. Furthermore, the residual mean square and all contrast sum of squares adjusted for the missing value covariates are their correct values. Although this method is quite attractive in some ways, it often cannot be implemented directly because specialized ANOVA routines may not have the capability to handle multiple covariates. It turns out, however, that Bartlett's method can be applied using only the complete-data ANOVA routine and a routine to invert an m×m symmetric matrix. Least square estimates of missing value by ANCOVA using only complete data method, correct least square estimates of standard errors and one degree of freedom sum of squares, correct least square sum of squares with more than one degree of freedom are some relevant least square estimates for estimating missing values in analysis of variance. There are several techniques to estimate missing values.

2. New Method and Methodology

Let there are $(n-1)$ observations and 1 missing observation. We want to estimate the missing observation. We know neither the missing value nor the distribution from where the observations are drawn. So, we know nothing about the parameters of the distribution or other characteristics like skewness, kurtosis, mean, median, mode, variance, higher order moments or even the tail behaviors of the distribution. In this situation we will estimate all the aforesaid characteristics and their volatility due to the change of sample size. We will also measure the deviation of the estimated characteristics from those of the missing value. So, we adjust our estimates of various characteristics due to the exact sample size and bandwidth of each of the characteristics. Later, all the estimated characteristics will be used to find out several relations among themselves to predict the probability distribution. The parameters will also be estimated under the predicted probability distribution. Later on, the deviation of the theoretically estimated characteristics and practically observed characteristics can be found to check how better the predicted distribution was by virtue of checking the equivalence of the theoretical and observed characteristics. Maximum Likelihood Function and the consistent rate of the mean sum of squares of error can be found to be confirmed that the performance of the estimated missing value and the error conducted due to the estimated missing value is least.

Let there are n observations out of which $(n-1)$ non-missing observations and one missing observation. Let the observations $x_{1,}x_{2,}\dots,\ x_{n-1}$ are non-missing and one observation x_n is missing. We want to estimate x_n. So out of $(n-1)$ non-missing observations, $(n-1)$ samples each of which is of size $(n-2)$ can be drawn assuming each sample has one missing observation. Assuming one non-missing observation as a missing one we can generate $(n-1)$ samples each of which is consisting of $(n-2)$ non-missing observations pretending the rest non-missing observations as the missing observation. So the $(n-1)$ generated samples are as below:

$(\boldsymbol{n-1})$ **samples each of size** $(\boldsymbol{n-2})$	**Assumed missing observation**
$x_{1,}x_{2,}\dots, x_{n-2}$	x_{n-1}
$x_{1,}x_{2,}\dots, x_{n-1}$	x_{n-2}
$\cdots$	$\cdots$
$x_{1,}x_{3,}\dots, x_{n-2}$	x_2
$x_{2,}x_{3,}\dots, x_{n-1}$	x_1

So we have calculated a class of characteristics (demonstrated in Table A1) to develop and observe several relationships among themselves (characteristics). For each of these characteristics, we will observe it's deviation from the same characteristic with the presence of dummy missing observation. Let us at first explain the easiest characteristic say sample mean and its deviation from the assumed missing value as addressed in Table A2.

Now, $$L = f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)\dots f(x_{n-1};\bar{x},S^2)$$

$$\log(L) = log[f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)\dots f(x_{n-1};\bar{x},S^2)]$$

$$\log(L) = log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2)) + \cdots + log(f(x_{n-1}; \bar{x}, S^2))$$

$$\therefore \frac{1}{n-1}\log(L) = \frac{1}{n-1}\sum_{i=1}^{n-1}\log(f(x_i; \bar{x}, S^2))$$

which can be termed as the average expected log likelihood function or expected log likelihood rate. Now, we should generate short incremented (various) values for x form the following range

$$\left(\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1}, \frac{1}{n-1}\sum_{i=1}^{n-1} x_i + \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1}\right).$$

Here *k* may be 0.50 or 1 or 2 or so on. The increment h can take the value 0.01 or 0.05 or 0.10 and so on. The values could be as below

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1},$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1} + h,$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1} + 2h,$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1} + 3h,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots,$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i + \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1}.$$

If we assume any of the aforesaid observations as the estimate of the *n*th pretended missing observation, and (if we consider) the available original observations $x_{1,}x_{2,} \dots, x_{n-1}$ as the $(n-1)$ other non-missing observations then the consecutive Maximum Likelihood Function or Likelihood Rate will be

$$L' = f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_n; \bar{x}, S^2)$$

$$\log(L') = log[f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_n; \bar{x}, S^2)]$$

$$\log(L') = log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2) + \cdots + log(f(x_n; \bar{x}, S^2))$$

$$\frac{1}{n}\log(L') = \frac{1}{n}\sum_{i=1}^{n}\log(f(x_i; \bar{x}, S^2))$$

We will search the incremented value of the nth observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\frac{1}{n-1}\log(L) = \frac{1}{n-1}\sum_{i=1}^{n-1}\log(f(x_i;\bar{x},S^2)) \cong \frac{1}{n}\log(L') = \frac{1}{n}\sum_{i=1}^{n}\log(f(x_i;\bar{x},S^2)).$$

The incremented value of the n^{th} observation for which the likelihood functions are same, will be an efficiently-estimated value of the missing observations.

However, if we get more than two estimates of the missing observation, we can check for which estimate of the missing value the first two moments are close to those of the original $(n-1)$ observations. Hence we will find the closer estimate of the missing value. Therefore, if we get more than two or three or more estimates of a missing observation, we can use all the estimates to estimate that missing value.

So, we have described how $(n-1)$ samples have been generated assuming one non-missing observation as a missing one in each case and calculated their sample averages to find out a bandwidth for the missing value. Here the missing value has been determint adding the half of the bandwidth of the missing value with the average of all of the available non-missing values. Similarly, several sample characteristics and their bandwidth can be calculated to find out different characteristics of the missing data as well as the distribution from which the sample (consisting of missing value and non-missing value) has been drawn. So, sample variance, sample higher order moments, sample median, mode, skewness, kurtosis, tail behaviors, etc. can be found using their respective bandwidth. Several relationships can be explored from the aforesaid estimated characteristics to recognize the pattern of the distribution and its relevant features. The relevant features, estimated parameters and the predicted distribution are used to fit the observed sample data. So least square fitting or least deviation fitting or any sort of other goodness of fit can be used to check the performance of the predicted probabilistic model along-with the bandwidth based estimated parameters and the characteristics. After checking the fitting performance of the predicted model for the observed data, we can observe whether the average log-likelihood function for both the non-missing and missing values is equivalent that of the average log-likelihood rate for the all non-missing values.

For more clarification let $n = 5$. So there are 4 non-missing observations and one missing observation. The non-missing observations are $x_{1,}x_{2,}x_3, x_4$ and the missing observation is x_5. So, assuming one non-missing observation as a missing one we can generate 4 samples each of which is consisting of 3 non-missing observations assuming the rest non-missing observations as the missing observation. So the 4 samples are as below:

Samples of size 3	Assumed missing observation
$x_{1,}x_{2,}x_3$	x_4
$x_{1,}x_{2,}x_4$	x_3
$x_{1,}x_{3,}x_4$	x_2
$x_{2,}x_{3,}x_4$	x_1

So we have calculated a class of characteristics (Table A3) to develop and observe some relationships among them (characteristics). For each of these characteristics we will observe it’s deviation from the same characteristic with the presence of assumed missing observation. Let us at first explain the easiest characteristics say sample mean and its deviation from the assumed missing value in the Table A4.

Now, $$L = f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)$$

$$\log(L) = log[f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)]$$

$$\log(L) = log(f(x_1;\bar{x},S^2)) + log(f(x_2;\bar{x},S^2)) + log(f(x_3;\bar{x},S^2)) + log(f(x_4;\bar{x},S^2))$$

$$\tfrac{1}{4}\log(L) = \tfrac{1}{4}\textstyle\sum_{i=1}^{4}\log(f(x_i;\bar{x},S^2))$$

which can termed as the average expected log likelihood or expected log likelihood rate. Now, we should generate short incremented various values form the range

$$\begin{pmatrix}\tfrac{1}{4}\sum_{i=1}^{4}x_i - \text{k}\frac{\overline{|x_1}-x_4|+\overline{|x_2}-x_{3|}+\overline{|x_3}-x_{2|}+\overline{|x_4}-x_1|}{4}, \\ \tfrac{1}{4}\sum_{i=1}^{4}x_i + \text{k}\frac{\overline{|x_1}-x_4|+\overline{|x_2}-x_{3|}+\overline{|x_3}-x_{2|}+\overline{|x_4}-x_{1|}}{4}\end{pmatrix}.$$

Here k may be 0.50 or 1 or 2 or so on. The increment h can take the value 0.01 or 0.05 or 0.10 and so on. The values the values could be

$$\tfrac{1}{4}\textstyle\sum_{i=1}^{4}x_i - \text{k}\frac{\overline{|x_1}-x_4|+\overline{|x_2}-x_{3|}+\overline{|x_3}-x_{2|}+\overline{|x_4}-x_1|}{4},$$

$$\tfrac{1}{4}\textstyle\sum_{i=1}^{4}x_i - \text{k}\frac{\overline{|x_1}-x_4|+\overline{|x_2}-x_{3|}+\overline{|x_3}-x_{2|}+\overline{|x_4}-x_1|}{4} + h,$$

$$\tfrac{1}{4}\textstyle\sum_{i=1}^{4}x_i - \text{k}\frac{\overline{|x_1}-x_4|+\overline{|x_2}-x_{3|}+\overline{|x_3}-x_{2|}+\overline{|x_4}-x_1|}{4} + 2h,$$

$$\tfrac{1}{4}\textstyle\sum_{i=1}^{4}x_i - \text{k}\frac{\overline{|x_1}-x_4|+\overline{|x_2}-x_{3|}+\overline{|x_3}-x_{2|}+\overline{|x_4}-x_1|}{4} + 3h,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,$$

$$\tfrac{1}{4}\textstyle\sum_{i=1}^{4}x_i + \text{k}\frac{\overline{|x_1}-x_4|+\overline{|x_2}-x_{3|}+\overline{|x_3}-x_{2|}+\overline{|x_4}-x_1|}{4}.$$

If we assume any of the afore said observations as the 5th observation and the four other observations are the given original observations $x_{1,}x_{2,}x_3,x_4$; then the consecutive maximum likelihood function or observed likelihood rate will be

$$L' = f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)\ f(x_5;\bar{x},S^2)$$

$$\log(L') = log[f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)f(x_5;\bar{x},S^2)]$$

$$\log(L') = log(f(x_1;\bar{x},S^2)) + log(f(x_2;\bar{x},S^2)) + log(f(x_3;\bar{x},S^2)) \\ + log(f(x_4;\bar{x},S^2)) + log(f(x_5;\bar{x},S^2))$$

$$\tfrac{1}{5}\log(L') = \tfrac{1}{5}\textstyle\sum_{i=1}^{5}\log(f(x_i;\bar{x},S^2))$$

We will search the incremented value of the 5th observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\tfrac{1}{4}\log(L) = \tfrac{1}{4}\textstyle\sum_{i=1}^{4}\log(f(x_i;\bar{x},S^2)) \cong \tfrac{1}{5}\log(L') = \tfrac{1}{5}\textstyle\sum_{i=1}^{5}\log(f(x_i;\bar{x},S^2)).$$

The incremented value of the 5th observation for which the likelihood functions are same, will be the estimated value of the missing observations. If we get more than two estimates of the missing observation (since we get two value of the 5th observation for whom the likelihood rates are same), we can check for which estimate of the missing value the first two moments are close to those of the original 4 observations. Hence we will find the estimate of the missing value.

3. Real Life Examples

We like to simulate a couple of samples each of which is of size n from a probability distribution with specified parameters. Later we will keep one observation a complete missing observation and pull it out from the original sample. Hence the original sample turns to a sample of size $n-1$. Out of $n-1$available observations of the sample, we will draw$n-1$ samples each of which is of size $n-2$. For each of the $n-1$ samples of size $n-2$, we will assume the absent observation as a dummy missing value of the sample. So, for each of the $n-1$ samples, there are $n-2$ available observations and one dummy missing value. From each of the $n-1$ samples, we will have one absolute dispersion between the average of $n-2$ available observations and the dummy missing observation. So, we will have $n-1$ absolute between differences for $n-1$ pairs of averages and dummy missing values. Averaging the $n-1$ absolute differences, we will calculate average absolute difference. Based on the average absolute difference, we will generate a possible range of the original missing value. We will generate several values of that range starting from the lower limit and will get several valued for fixed increment upto to upper limit of that range. We will check whether the average log likelihood of the $n-1$original observations is similar for which n^{th} observed missing value from the generating range and the $n-1$ observations.

Example 3.1

Let $n=10$ So there are 9 non-missing observations and one missing observation. The non-missing observations are 1.729466, 3.547037, 3.6597, 5.814905, 3.817457, 6.333606, 4.05684, 3.748781, 3.608116 and the missing observation is2.671239. The average of this nine non-missing observations are 4.0351. Now, assuming one non-missing observation as a missing one we can generate 9 samples each of which is consisting of 8 non-missing observations assuming the rest non-missing observations as the missing observation. So the 9 sampleseach consisting of 8 non-missing values are as below (the bold numbers in the last row are representing here the assumed missing value for each sample):

1.73	1.73	1.73	1.73	1.73	1.73	1.73	1.73	3.61
3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.55	3.55
3.55	3.55	3.55	3.55	3.55	3.55	3.66	3.66	3.66
3.66	3.66	3.66	3.66	3.66	5.81	5.81	5.81	5.81
5.81	5.81	5.81	5.81	3.82	3.82	3.82	3.82	3.82
3.82	3.82	3.82	6.33	6.33	6.33	6.33	6.33	6.33
6.33	6.33	4.06	4.06	4.06	4.06	4.06	4.06	4.06
4.06	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75
3.75	***4.06***	***6.33***	***3.82***	***5.81***	***3.66***	***3.55***	***3.61***	***1.73***

After using Table A5 and table A6, we obtain the Expected Log Likelihood Rate is 0.720. By using the formula shown above, we get the range as (2.4976, 5.5726); where k=1. Let the increment, h=0.1. For each increment we will get average log likelihood rate for 10 observations. And for the third increment (incremented value=2.7976) we get the same value for the Expected Average Log Likelihood and Observed Average Log Likelihood. So, our estimated value of the missing observation is 2.7976

Example 3.2

Let, $n = 10$.So there are 9 non-missing observations and one missing observation. The non-missing observations are 1.729466, 3.547037, 3.6597, 5.814905, 3.817457, 6.333606, 3.748781, 2.671239, 3.608116 and the missing observation is 4.05684. The average of these nine non-missing observations are 3.8811. Now, assuming one non-missing observation as a missing one we can generate 9 samples each of which are consisting of 8 non-missing observations assuming the rest non-missing observation as the missing observation. So, 9 samples each consisting of 8 non-missing values are given below (the bold red colored numbers in the last row are representing here the assumed missing value for each sample). The range is (2.2363,5.5260); where, k=1 and increment, h= 0.05

1.73	1.73	1.73	1.73	1.73	1.73	1.73	1.73	3.61
3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.55	3.55
3.55	3.55	3.55	3.55	3.55	3.55	3.66	3.66	3.66
3.66	3.66	3.66	3.66	3.66	5.81	5.81	5.81	5.81
5.81	5.81	5.81	5.81	3.82	3.82	3.82	3.82	3.82
3.82	3.82	3.82	6.33	6.33	6.33	6.33	6.33	6.33
6.33	6.33	3.75	3.75	3.75	3.75	3.75	3.75	3.75
3.75	2.67	2.67	2.67	2.67	2.67	2.67	2.67	2.67
2.67	***3.75***	***6.33***	***3.82***	***5.81***	***3.66***	***3.55***	***3.61***	***1.73***

Using the calculations of Table A7 and Table A8, we get the Observed Average Log Likelihood Rate -0.744. For the incremented values 2.5363 and 5.2363 the value of the Expected Average Log Likelihood and Observed Average Log Likelihood are same. The average of these two values are 3.8863. Therefore, our estimated missing value is 3.8863

Conclusion

The missing technique is a kind of check and balance method in estimating the missing value. In each step it checks the fluctuation due to sample size and balance it by capturing the dispersion of the estimate of the known data from the assumed unknown data which is really known. So this method is trying to find the original rate of change of the deviation from the missing value for the exact size of the realized sample. So from two directions, one direction from sample size and other direction for the deviation from the missing value, the missing technique has been aided to estimate the missing value efficiently maintaining a good performance through several goodness of fit tests. We will provide later the extended version of the estimation of more than one missing value in the sample in this paper.

Reference

Afifi, A. A., & Elashoff, R. M. (1966). Missing observations in multivariate statistics I. Journal of American Statistical Association, 61 (315) 595-605.

Allan, F. G. and Wishart, J. (1930). A method of estimating the yield of a missing plot in field experiments. J. Agric. Sci. 20, 399-406.

Bartlett, M. S. (1937). Some examples of statistical methods of reserahc in agriculture and applied botany. J. Roy. Statis. Soc. B4. 137-170.

Dempster, Laird, and Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society, series B.

Hartley. H. O.(1958). Maximum likelihood estimation from incomplete data. Biometrcis. 14, 174-194.

Hartley, H. O. and Hocking, R. R. (1971). The Analysis of Incomplete Data. Biometrics, 27, 783-823.

Healy, M J. R. and Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers, Appl. Statis.. 5, 203-206.

Little. R. J. A, Rubin. D. B. (1983). Incomplete data, Encyclopedia of the Statistical Sciences. 4, 46-53.

Little, R. J. A. and Schenker, N. (1994). Missing Data, in Handbook for Statistical Modeling in the Social and Behavioral sciences. 39-75. New York: Plenum.

Little, R. J. A. (1997). Bio statistical Analysis with missing data, in Encyclopedia of Biostatistics, London: Wiley.

Little, R. J. A, Rubin. D. B. (2002). Statistical Analysis with Missing Data. 2nd edition. Wiley Publishers.

Nie, et al (1975). SPSS, 2nd ed. New York: McGraw Hill.

Orchard and Woodbury (1972). A missing information principle: theory and applications. Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1972), 697-715.

Pearce, S. C. (1965). Biological Statistics: An Introduction. New York:McGraw-Hill.

Wilkinson, G. N. (1958). Estimation of missing values for the analysis of incomplete data. Biometrics. 14, 257-286.

Appendix

Table A1: Mean and variance table for $(n-1)$ samples

	Sample Mean	**Sample Variance**
	$\overline{x_1} = \frac{x_{1+}x_{2+} \dots + x_{n-2}}{n-2}$	$S_1^2 = \frac{(x_1 - \overline{x_1})^2 + (x_2 - \overline{x_1})^2 + (x_{n-2} - \overline{x_1})^2}{n-3}$
	$\overline{x_2} = \frac{x_{1+}x_{2+\cdots+}x_{n-1}}{n-2}$	$S_2^2 = \frac{(x_1 - \overline{x_2})^2 + (x_2 - \overline{x_2})^2 + (x_{n-1} - \overline{x_2})^2}{n-3}$
	…	…
	$\overline{x_{n-2}} = \frac{x_{1+}x_{3+\cdots+}x_{n-2}}{n-2}$	$S_{n-2}^2 = \frac{(x_1 - \overline{x_{n-2}})^2 + (x_3 - \overline{x_{n-2}})^2 + \cdots + (x_{n-2} - \overline{x_{n-2}})^2}{n-3}$
	$\overline{x_{n-1}} = \frac{x_{2+}x_{3+\cdots+}x_{n-1}}{n-2}$	$S_{n-1}^2 = \frac{(x_2 - \overline{x_{n-1}})^2 + (x_3 - \overline{x_{n-1}})^2 + \cdots + (x_{n-1} - \overline{x_{n-1}})^2}{n-3}$
Average	$\bar{x} = \frac{\overline{x_1} + \overline{x_2} + \cdots + \overline{x_{n-1}}}{n-1}$	$S^2 = \frac{S_1^2 + S_2^2 + \cdots + S_{n-1}^2}{n-1}$

Table A2: Difference table for $(n-1)$ samples

$(\boldsymbol{n-1})$ **sample means each of size** $(\boldsymbol{n-2})$	**Assumed Missing Value**	**Difference**	**|Difference|**
$\overline{x_1} = \frac{x_{1+}x_{2+} \dots + x_{n-2}}{n-2}$	x_{n-1}	$\overline{x_1} - x_{n-1}$	$\lvert\overline{x_1} - x_{n-1}\rvert$
$\overline{x_2} = \frac{x_{1+}x_{2+\cdots+}x_{n-1}}{n-2}$	x_{n-2}	$\overline{x_2} - x_{n-2}$	$\lvert\overline{x_2} - x_{n-2}\rvert$
…	…	…	…
$\overline{x_{n-2}} = \frac{x_{1+}x_{3+\cdots+}x_{n-2}}{n-2}$	x_2	$\overline{x_{n-2}} - x_2$	$\lvert\overline{x_{n-2}} - x_2\rvert$
$\overline{x_{n-1}} = \frac{x_{2+}x_{3+\cdots+}x_{n-1}}{n-2}$	x_1	$\overline{x_{n-1}} - x_1$	$\lvert\overline{x_{n-1}} - x_1\rvert$
Total			$\lvert\overline{x_1} - x_{n-1}\rvert + \lvert\overline{x_2} - x_{n-2}\rvert + \cdots + \lvert\overline{x_{n-2}} - x_2\rvert + \lvert\overline{x_{n-1}} - x_1\rvert$
Average			$\frac{\lvert\overline{x_1} - x_{n-1}\rvert + \lvert\overline{x_2} - x_{n-2}\rvert + \cdots + \lvert\overline{x_{n-2}} - x_2\rvert + \lvert\overline{x_{n-1}} - x_1\rvert}{n-1}$

Table A3: Mean and variance table for 4 samples

	Sample Mean	**Sample Variance**
	$\overline{x_1} = \frac{x_{1+}x_{2+}x_3}{3}$	$S_1^2 = \frac{(x_1 - \overline{x_1})^2 + (x_2 - \overline{x_1})^2 + (x_3 - \overline{x_1})^2}{3-1}$
	$\overline{x_2} = \frac{x_{1+}x_{2+}x_4}{3}$	$S_2^2 = \frac{(x_1 - \overline{x_2})^2 + (x_2 - \overline{x_2})^2 + (x_4 - \overline{x_2})^2}{3-1}$
	$\overline{x_3} = \frac{x_{1+}x_{3+}x_4}{3}$	$S_3^2 = \frac{(x_1 - \overline{x_3})^2 + (x_3 - \overline{x_3})^2 + (x_4 - \overline{x_3})^2}{3-1}$
	$\overline{x_4} = \frac{x_{2+}x_{3+}x_4}{3}$	$S_4^2 = \frac{(x_2 - \overline{x_4})^2 + (x_3 - \overline{x_4})^2 + (x_4 - \overline{x_4})^2}{3-1}$
Average	$\bar{x} = \frac{\overline{x_1} + \overline{x_2} + \overline{x_3} + \overline{x_4}}{4}$	$S^2 = \frac{S_1^2 + S_2^2 + S_3^2 + S_4^2}{4}$

Table A4: Difference table for 4 samples

Sample Mean of size 3	**Assumed Missing Value**	**Difference**	$\|\boldsymbol{Difference}\|$
$\overline{x_1} = \frac{x_{1+}x_{2+}x_3}{3}$	x_4	$\overline{x_1} - x_4$	$\|\overline{x_1} - x_4\|$
$\overline{x_2} = \frac{x_{1+}x_{2+}x_4}{3}$	x_3	$\overline{x_2} - x_3$	$\|\overline{x_2} - x_3\|$
$\overline{x_3} = \frac{x_{1+}x_{3+}x_4}{3}$	x_2	$\overline{x_3} - x_2$	$\|\overline{x_3} - x_2\|$
$\overline{x_4} = \frac{x_{2+}x_{3+}x_4}{3}$	x_1	$\overline{x_4} - x_1$	$\|\overline{x_4} - x_1\|$
Total			$\|\overline{x_1} - x_4\| + \|\overline{x_2} - x_3\|$ $+\|\overline{x_3} - x_{2\|} + \|\overline{x_4} - x_1\|$
Average			$\frac{\|\overline{x_1} - x_4\| + \|\overline{x_2} - x_3\| + \|\overline{x_3} - x_{2\|} + \|\overline{x_4} - x_1\|}{4}$

Table A5: Mean and variance table for 9 samples

	Sample Mean	**Sample Standard Deviation**
	4.07	1.43
	4.03	1.44
	3.75	1.10
	4.06	1.43
	3.81	1.25
	4.08	1.43
	4.10	1.42
	4.09	1.43
	4.32	1.10
Average	**4.04**	**1.34**

Table A6: Difference table for 9 samples.

Sample mean of size 8	**Assumed Missing Value**	**Absolute Difference**
4.07	3.75	0.48
4.03	4.06	0.04
3.75	6.33	3.88
4.06	3.82	0.37
3.81	5.81	3.00
4.08	3.66	0.63
4.10	3.55	0.82
4.09	3.61	0.72
4.32	1.3	3.89
Total		**13.73**
Average		**1.54**

Table A7: Mean and variance table for 9 samples.

	Sample Mean	**Sample Standard Deviation**
	4.03	1.44
	3.90	1.52
	3.57	1.15
	3.89	1.52
	3.64	1.30
	3.91	1.51
	3.92	1.51
	3.92	1.51
	4.15	1.25
Average	**3.88**	**1.41**

Table A8: Difference table for 9 samples.

Sample Mean of Size 8	**Assumed Missing Value**	**Absolute Difference**
4.03	2.67	2.04
3.90	3.75	0.22
3.57	6.33	4.14
3.89	3.82	0.11
3.64	5.81	3.26
3.91	3.66	0.37
3.92	3.55	0.56
3.92	3.1	0.46
4.15	1.73	3.63
Total		**14.79**
Average		**1.64**