International Journal of Academic and Applied Research (IJAAR)
ISSN: 2000-005X
Vol. 2 Issue 10, October – 2018, Pages: 15-22

Comparing Methods of Estimating Missing Values in One-Way Analysis of Variance

[*]Chukwunenye, Victor Gozie, Eze, Francis Chkwuemeka

Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
[*]Email: gozievictor2012@gmail.com

***Abstract*:** *It is obvious that the treatment of missing data has been an issue in statistics for some time now, and hence has started gaining the attention of researchers. This paper established the various methods usable in estimating missing values, determined which of the methods is the best in estimating missing values in one–way analysis of variance (ANOVA), determined at which percentage level of Missingness is the method best and verified the effect of missing values on the statistical power and non–centrality parameters in one–way ANOVA. The methods examined are Pairwise Deletion (PD), Mean Substitution (MS), Regression Estimation (RE), Multiple Imputation (MI) and Expectation Maximization (EM). Mean Square Errors (MSEs), that is variances of the methods were compared. It was found that MS had the least variance at 5, 10, 15, and 25 percent levels of Missingness while EM had the least variance at 20 percent Missingness level. PD method yielded the least statistical power at all the percentage levels of Missingness. Non–centrality parameters increased with increasing percentage level of Missingness and it was also found that at 25 percent level of Missingness (after 20 percent), the statistical power started to reduce. EM method was recommended since MS yielded the least MSEs because of its limitations. Meanwhile PD should not be an option while dealing with missing data in one – way ANOVA due to loss of statistical power and possibly increased MSE.*

Keywords: Pairwise deletion; Mean Substitution; Regression Estimation; Multiple Imputation; Expectation Maximization.

1. INTRODUCTION

Missing data are common problem facing researchers. There are some reasons why data are missing. These may include ignoring values in datasets by respondents refusing to respond to questionnaires. In some cases, high data collection may as well cause missing data. A wild value such as age being recorded as negative could be regarded as missing data.

Missing data can introduce ambiguity into data analysis. Working with missing data can affect properties of statistical estimators such as means and variances, resulting in a loss or reduction of statistical power and committing either type 1 or type 2 error. To avoid these problems, researchers are faced with two options: (a) to delete those cases which have missing data, or (b) to fill-in the missing values with estimated values, (Acock 2005; Howell 2008; Schmitt et al 2015; Tanguma 2000). In missing data, common statistical method of analysis becomes inappropriate and difficult to apply. In a case where data are missing in a factorial analysis of variance, the design is said to be unbalanced and the appropriate standard statistical analysis can no longer apply. Even if data are assumed to be missing in a completely random fashion, the proper analysis is completely complicated, (Jain et al 2016: Peng et al 2003).

With the advent of computer software, sophisticated analyses of missing data can now be accomplished. Best practices related to missing data in research call for two items of essential information that should be reported in every research study: (a) the extent and nature of missing data and (b) the procedures used to manage the missing data, including the rationale for using the method selected, (Schlomer et al 2010).

Dealing with unequal sample size in analysis of variance (ANOVA), the F-statistic will be more sensitive to small departure from the assumption of equal variance (homoscedasticity) compared to the equal sample size treatment analysis. If the homoscedasticity assumption is violated, then the treatment effect produced by ANOVA will be a biased one.

Many researchers have proposed several methods for dealing with missing data. Example of such methods is Listwise Deletion, which decreases the number of observations further and can result in biased results when applied to small data set (Sikka et al 2010). Mean substitution is another method which has some limitations according to Cool (2000) and Little and Rubin (1987): (a) sample size is overestimated, (b) correlations are negatively biased, (c) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean and (d) variance is underestimated.

Myrtveit et al (2001) evaluated four missing data techniques (MDTs) in the context of software cost modelling. The techniques evaluated are Listwise Deletion (LD), Mean Imputation (MI), Similar Response Pattern Imputation (SRPI) and Full Information Maximum Likelihood (FIML). They applied the MDTs to a data set and constructed a regression-based prediction models. Their evaluation suggests that only FIML is appropriate when the data are not missing completely at random (MCAR).

A similar work was done by Tanguma (2000) where he worked on four methods of dealing with missing data. Four commonly used methods namely: listwise deletion, pairwise deletion, mean imputation and regression imputation were considered using hypothetical data set. In his work, listwise deletion, which is the default in some statistical packages (e.g., the Statistical Package for the Social Sciences and the Statistical Analysis System), is the most commonly used method. He claimed that listwise deletion eliminates all cases for a participant missing data on any predictor or criterion variable, it is not the most effective method. Pairwise deletion uses those observations that have no missing values to compute the correlations. Thus, it preserves information that would have been lost when using listwise deletion. In mean imputation, the mean for a particular variable, computed from available cases, is substituted in place of missing data values on the remaining cases. This allows the researcher to use the rest of the participant's data. The researcher found that when using a regression-based procedure to estimate the missing values, the estimation takes into account the relationships among the variables. Thus, substitution by regression is more statistically efficient he concluded.

Song et al (2005) noted that selecting the appropriate imputation technique can be a difficult problem. One reason for this being that the techniques make assumptions about the underlying missingness mechanism; that is how the missing values are distributed within the data set. It is compounded by the fact that, for small data sets, it may be very difficult to determine what is the missingness mechanism. This means there is a danger of using an inappropriate imputation technique. They therefore said that it is necessary to determine what is the safest default assumption about the missingness mechanism for imputation techniques when dealing with small data sets. The research was done with two simple and commonly used techniques: Class Mean Imputation (CMI) and k Nearest Neighbors (k-NN) coupled with two missingness mechanisms: missing completely at random (MCAR) and missing at random (MAR). They had two conclusions. They concluded that for their analysis CMI is the preferred technique since it is more accurate and more importantly, the impact of missingness mechanism on imputation accuracy is not statistically significant. This is a useful finding since it suggests that even for small data sets we can reasonably make a weaker assumption that the missingness mechanism is MAR. Thus both imputation techniques have practical application for small software engineering data sets with missing values.

Horton and Kleinman (2007) worked on a comparison of missing data methods and software to fit incomplete data regression models. They highlighted that missing data are a recurring problem that can cause bias or lead to inefficient analyses, noting that each of the approaches to dealing with missing data is more complicated when there are many patterns of missing values, or when both categorical and continuous random variables are involved. They noted that implementations of routines to incorporate observations with incomplete variables in regression models are now widely available. They reviewed the routines in the context of a motivating example from a large health services research dataset. While there are still limitations to the current implementations, and additional efforts are required of the analyst, they advised that it is feasible to incorporate partially observed values, and those methods should be used in practice.

Twala et al (2006) worked on ensemble of missing data techniques to improve software prediction accuracy saying that software engineers are commonly faced with the problem of incomplete data. They also said that incomplete data can reduce system performance in terms of predictive accuracy. It was however noted that unfortunately, rare research has been conducted to systematically explore the impact of missing values, especially from the missing data handling point of view as regards software prediction accuracy. This has made various missing data techniques (MDTs) less significant. Their paper described a systematic comparison of seven MDTs using eight industrial datasets. Their findings from an empirical evaluation suggest listwise deletion as the least effective technique for handling incomplete data while multiple imputation achieves the highest accuracy rates. They further proposed and showed how a combination of MDTs by randomizing a decision tree building algorithm leads to a significant improvement in prediction performance for missing values up to 50%.

Cool (2000) reviewed of methods for dealing with missing data reviewing some of the various strategies for addressing the missing data problem. The research showed that which technique to use best depends on several factors. The paper opined that listwise deletion and pairwise deletion methods both result in a reduction in sample size which leads to reduced precision in the estimates of the population parameters. This reduction in sample size also reduces the power of statistical significance testing, and this poses a potential threat to statistical conclusion validity. Although the same attenuation of the correlation coefficicient occur, the methods of inserting means and using regression analyses are about equally effective under conditions of low multicollinearity the paper argued. The most important advantages of these mean imputation methods are the retention of sample size and, consequently of statistical power in subsequent analyses. She noted that unfortunately, because of the numerous factors influencing the relative success of the competing techniques, no one method for handling the missing data problem has been shown to be uniformly superior.

Saunders et al (2006) compared methods of imputing missing data for social work researchers noting that choosing the most appropriate method to handle missing data during analyses is one of the most challenging decisions confronting researchers. In their work, six methods of data imputation were used to replace missing data from two data sets of varying sizes and the results were examined. The methods

used are listwise deletion, mean substitution, hotdecking, regression imputation or conditional mean imputation, single implicate and multiple implicate. Each imputation method was defined, and the pros and cons of its use in social science research are identified. They discussed comparisons of descriptive measures and multivariate analyses with the imputed variables and the results of a timed study to determine how long it took to use each imputation method on first and subsequent use. The results of the statistical analysis conducted for their study suggest that a large sample with only a small percentage of missing values is not influenced to the same degree by data imputation methods as are smaller data sets. They said however, that regardless of the sample size, researchers should still consider the advantages and disadvantages in choosing the most appropriate imputation method. In conclusion, they added that every researcher should explore the patterns of missing values in data set and consider constructing instruments to clearly identify some patterns of missingness; since social work can no longer avoid the issues of missing data, every research report should report the reasons for and the amount of missing data as well as what data imputation method was used during the analysis; multiple implicate is currently the best imputation method and should be used whenever possible.

Eekhout et al (2012) did a systematic review of how missing values are reported and handled, with **t**he objectives of examining how researchers report missing data in questionnaires and to provide an overview of current methods for dealing with missing data. They included 262 studies published in 2010 in three leading epidemiological journals. Information was extracted on how missing data were reported, types of missing, and methods for dealing with missing data. They discovered that 78% of studies lacked clear information about the measurement instruments; missing data in multi–item instruments were not handled differently from other missing data; Complete-case analysis was most frequently reported (81% of the studies), and the selectivity of missing data was seldom examined. They noted that although there are specific methods for handling missing data in item scores and in total scores of multi – item instruments, these are seldom applied. Researchers mainly use complete – case analysis for both types of missing data, which may seriously bias the study results.

Xu (2001) investigated properties and effects of three selected missing data handling techniques (listwise deletion, hot deck imputation, and multiple imputation) via a simulation study, and applied the three methods to address the missing race problem in a real data set extracted from the National Hospital Discharge Survey. The results of the study showed that multiple imputation and hot deck imputation procedures provided more reliable parameter estimates than listwise deletion. A similar outcome was observed with respect to the standard errors of the parameter estimates, with the multiple imputation and hot deck imputation producing parameter estimates with smaller standard errors. Multiple imputation outperformed the hot deck imputation by using larger significant levels for variables with missing data and reflecting the uncertainty with missing values. In summary, the study showed that employing an appropriate imputation technique to handling missing data in public use surveys is better than ignoring the missing data.

Myers (2011) did a research titled 'Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data'. The paper revealed that even though missing data are a ubiquitous problem in quantitative communication research, yet the missing data handling practices found in most published work in communication leave much room for improvement. In the article, problems with current practices were discussed and suggestions for improvement were offered. Finally, hot deck imputation was suggested as a practical solution to many missing data problems. A computational tool for SPSS (Statistical Package for the Social Sciences) was presented that will enable communication researchers to easily implement hot deck imputation in their own analyses.

Considering the ambiguity, bias and reduction in values of computed statistics (such as mean, variance, standard deviation, etc.) which arise as a result of missing values especially in one-way ANOVA, there is the need to embark on a research capable of coming up with the best method that can be recommended for use when there are missing values in one-way ANOVA.

2.0 Methodology

2.1 Investigated Techniques

Five missing value (MV) imputation techniques were investigated. They are Pairwise Deletion (PD), Mean Substitution (MS), Multiple Imputation (MI), Regression Imputation (RI) and Expected Maximization (EM).

2.1.1 Pairwise Deletion

According to Acock (2005), pairwise deletion uses all available information in the sense that all participants who answered a pair of variables are used regardless of whether they answered other variables.

He noted that one reason pairwise deletion is unpopular is that it can produce a covariance matrix that is impossible for any single sample. Specifically, because each covariance could be based on a different subsample of participants, the covariance does not have the constraints they would have if all covariance were based on the same set of participants. It is possible that the pairwise correlation matrix cannot be inverted, a necessary step for estimating the regression equation and structural equation models. This problem may appear in the program output as a warning that a matrix is not positive definite. This problem can occur even when the data meet the assumption of MCAR.

With pairwise deletion it is difficult to compute the degrees of freedom because different parts of the model have different samples. Selecting the sample size using the correlation that has the most observations would be a mistake and would exaggerate statistical power. Selecting

the sample size using the correlation that has the fewest observations would reduce power.

2.1.2 Mean Substitution

In this method, the missing data of an attribute is found by calculating mean of total values of that attribute. It assumes that a missing value for an individual on a given variable is best estimated by the mean (expected value) for the non-missing observations for that variable, Aruguma (2015), Cool (2000).

However, Cool (2000) listed some of the limitations of MS according to Little and Rubin (1987). The limitations are:

(a) sample size is overestimated,
(b) correlations are negatively biased,
(c) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean and
(d) variance is underestimated.

Acock (2015) argued that Mean Substitution is especially problematic when there are many missing values. For example, if 30% of the people do not report their income and $45.219 is substituted for each of them, then 30% of the sample has zero variance on income, thus greatly attenuating the variance of income. This attenuated variance leads to underestimating the correlation of income with any other variable.

For any variable *i* with *m* missing data in the ij[th] cell, MS imputes an estimate from the statistic:

$$\overline{Y}_{ijk} = \frac{1}{n-m}\sum_{k}^{n-m} y_{ijk} \qquad (2.1)$$

2.1.3 Multiple Imputation (MI)

Rubin (1976) gave a definition of MI to be a technique for handling survey non–response that replaces each missing value created by non–response by a vector of possible values that reflect uncertainty about which values to impute. Multiple imputation replaces each missing value by a vector composed of $M \geq 2$ possible values. The *M* values are ordered in the sense that the first components of the vectors for the missing values are used to create one completed data set; the second components of the vectors are used to create second completed data set and so on.

according to rubin (1986), the imputation task begins by sorting the sampled units by their pattern of missing data. index these units by $\boldsymbol{j = 0, 1, 2, \dots, J}$, where $\boldsymbol{j = 0}$ refers units with no missing data. the phenomenological bayesian framework tells us that, in general, each pattern of missing data suppose that each of z is modeled as independently and identically distributed (i.i.d.); $\boldsymbol{f(Z|\Phi_j)}$, $\boldsymbol{j = 0, 1, 2, \dots, J}$, where $\boldsymbol{\Phi_j = \Phi_j(\Phi)}$ and $\boldsymbol{\Phi}$ has posterior distribution pos($\boldsymbol{\Phi}$). when mechanisms are ignorable, $\boldsymbol{\Phi_j = \Phi_1 = \Phi_2 = \cdots = \Phi_j}$, and thus the rows of z are not only independent, they are also identically distributed.

for the j[th] pattern of missing data, z is partitioned into $\mathbf{Z} = (\mathbf{V_j}, \mathbf{U_j})$, where $\mathbf{V_j}$ are the missing variables and $\mathbf{U_j}$ are the observed variables; for $\mathbf{j = 0}$, $\mathbf{Z = U_0}$. since for each unit, we must impute values for $\mathbf{V_j}$ given the model and observed values of $\mathbf{U_j}$, we factor the density $\boldsymbol{f(Z|\Phi_j)}$ as:

$$\boldsymbol{f(Z|\Phi_j) = f(V_j|U_j, \xi_j) f(U_j, n_k)} \qquad (2.2)$$

Where $\boldsymbol{\xi_j = q_j(\Phi_j)}$ and $\boldsymbol{n_j = \bar{q}_j(\Phi_j)}$ and where $\mathbf{q_j}(.)$ and $\bar{\mathbf{q}}_\mathbf{j}(.)$ are the appropriate functions of the parameter $\boldsymbol{\Phi_j}$ corresponding to the partition $\mathbf{Z} = (\mathbf{V_j}, \mathbf{U_j})$.

Having noted the above, the imputation task is as follows:

Draw $\boldsymbol{\Phi_0}$ from the posterior distribution of $\boldsymbol{\Phi_0}$, pos($\boldsymbol{\Phi_0}$). Call the drawn value $\boldsymbol{\Phi_0^*}$.

For $\boldsymbol{j = 0, 1, 2, \dots, J}$

$$\begin{cases} \textbf{(i) draw } \boldsymbol{\Phi_j} \textbf{ from pos}(\boldsymbol{\Phi_j|\Phi_0 = \Phi_0^*, \dots, \Phi_{j-1} = \Phi_{j-1}^*}) \\ \textbf{(ii) calculate } \boldsymbol{\xi_j^* = q_j(\Phi_j^*)} \end{cases} \qquad (2.3)$$

2.1.4 Regression Estimation (RE)

Many of our problems, as well as many of the solutions that have been suggested concerning the use of RE, refer to designs that can roughly be characterized as linear regression models (Howel 2008).

Suppose that we have collected data on several variables. One or more of those variables is likely to be considered a dependent variable, and the others are predictor, or independent, variables. Our interest is that for a variable with missing value, we fit a model of the form:

$$\boldsymbol{Y_j = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \cdots + \beta_{(j-1)} Y_{(j-1)}} \qquad (2.4)$$

The fitted regression model has regression parameter estimates $(\boldsymbol{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{(j-1)}})$ and the associated covariance matrix $\boldsymbol{\sigma_j^2 V_j}$, where $\mathbf{V_j}$ is the usual $\mathbf{X'X}$ matrix from the intercept and variables $\mathbf{Y_1, Y_2, \dots, Y_{(j-1)}}$.

For each imputation, new parameters $(\boldsymbol{\beta_{*0}, \beta_{*1}, \dots, \beta_{*(j-1)}})$ and $\boldsymbol{\sigma_{*j}^2}$ are drawn from a posterior predictive distribution of the missing data. That is, they are simulated from $(\boldsymbol{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{(j-1)}})$, $\boldsymbol{\sigma_j^2}$ and $\mathbf{V_j}$.

The missing values are then replaced by

$$\boldsymbol{Y_j = \beta_{*0} + \beta_{*1} y_1 + \beta_{*2} y_2 + \cdots + \beta_{*(j-1)} y_{(j-1)} + z_i \sigma_{*j}} \qquad 3.8$$

Where $\boldsymbol{y_1, y_2, \dots, y_{(j-1)}}$ are the covariates of the first *(j – 1)* variables and $\boldsymbol{z_i}$ is a simulated normal deviate.

2.1.5 Expectation Maximization (EM)

EM is a general approach to iterative computation of maximum–likelihood estimates when the observations can be viewed as incomplete data. It is called EM algorithm because each iteration of the algorithm consists of an expectation step followed by a maximization step.

Dempster et al., (1977) postulated a family of sampling densities $f(x|\Phi)$ depending on parameter Φ and derive its corresponding family of sampling densities $g(y|\Phi)$. The

complete–data specification $f(...|...)$ is related to the incomplete–data specification $g(...|...)$ by

$$g(y|\Phi) = \int_{x(y)} f(x|\Phi)dx \qquad (2.5)$$

The EM algorithm is directed at finding a value of Φ which maximizes $g(y|\Phi)$ given an observed y, but it does so by making essential use of the associated family $f(x|\Phi)$.

2.2 Performance Indicators

The imputation techniques will be accessed using the variance (MSE) obtained from the one – way ANOVA of the data. This will be obtained when one – way ANOVA is performed using each set of data. The various imputation techniques will be also accessed with the use of the statistical power.

2.3 Method of Data Collection

Data for this research work will be obtained using data simulation. Simulated data will be tested to make sure they meet the assumptions of ANOVA. Real life data will as well be analyzed to compare the results.

3.0 Data Analysis

The test of assumption of ANOVA especially the normality and homogeneity tests for the simulated complete data and for the various methods of imputation for 5%, 10%, 15%, 20% and 25% missing data were performed using Shapiro-Wilk test Levene Statistic. The results show that the data are normally distributed and have constant variance.

Similarly, normality and homogeneity tests were performed on the real data obtained from Neter et al (1996) and the results show that the data are normally distributed and have constant variance.

3.1 Presentation of Mean Square Errors of Simulated Data

Here, results of the simulated data will be presented; the presentations will be done using tables 3.1.1 to 3.1.5.

Table 3.1.1: MSEs at 5 Percent Missing Level

S/No	MV Estimation Method	Normality	Homogeneity	MSE
1	Pair Wise Deletion	Normal	Homogeneous	7.605
2	Mean Substitution	Normal	Homogeneous	7.205
3	Multiple Imputation	Normal	Homogeneous	7.544
4	Regression Estimation	Normal	Homogeneous	7.423
5	Expectation Maximization	Normal	Homogeneous	7.388

Table 3.1.2: MSEs at 10 Percent Missing Level

S/No	MV Estimation Method	Normality	Homogeneity	MSE
1	Pair Wise Deletion	Normal	Homogeneous	7.945
2	Mean Substitution	Normal	Homogeneous	7.111
3	Multiple Imputation	Normal	Homogeneous	7.331
4	Regression Estimation	Normal	Homogeneous	9.016
5	Expectation Maximization	Normal	Homogeneous	7.133

Table 3.1.3: MSEs at 15 Percent Missing Level

S/No	MV Estimation Method	Normality	Homogeneity	MSE
1	Pair Wise Deletion	Normal	Homogeneous	7.609
2	Mean Substitution	Normal	Homogeneous	6.407
3	Multiple Imputation	Normal	Homogeneous	6.606
4	Regression Estimation	Normal	Homogeneous	7.352
5	Expectation Maximization	Normal	Homogeneous	7.078

Table 3.1.4: MSEs at 20 Percent Missing Level

S/No	MV Estimation Method	Normality	Homogeneity	MSE
1	Pair Wise Deletion	Normal	Homogeneous	6.6656
2	Mean Substitution	Normal	Homogeneous	5.468
3	Multiple Imputation	Normal	Homogeneous	5.693
4	Regression Estimation	Normal	Homogeneous	6.188
5	Expectation Maximization	Normal	Homogeneous	5.405

Table 3.1.5: MSEs at 25 Percent Missing Level

S/No	MV Estimation Method	Normality	Homogeneity	MSE
1	Pair Wise Deletion	Normal	Homogeneous	8.199
2	Mean Substitution	Normal	Homogeneous	6.042

3	Multiple Imputation	Normal	Homogeneous	6.855
4	Regression Estimation	Normal	Homogeneous	8.526
5	Expectation Maximization	Normal	Homogeneous	6.366

3.2 Presentation of Mean Square Errors of Data from Neter et al (1996)

Neter et al (1996) presented incomplete data as problem to be solved by readers of the book Applied Linear Statistical Models on productivity improvement (PI) and rehabilitation therapy (RT) and the result are presented in tables 3.2.1 and 3.2.2.

Table 3.2.1: MSEs of ANOVA results of Productivity Improvement

S/No	MV Estimation Method	MSE
1	Pair Wise Deletion	0.640
2	Mean Substitution	0.466
3	Multiple Imputation	0.642
4	Regression Estimation	0.773
5	Expectation Maximization	0.492

Table 3.2.2: MSEs of ANOVA results of Rehabilitation Therapy

S/No	MV Estimation Method	MSE
1	Pair Wise Deletion	19.810
2	Mean Substitution	15.407
3	Multiple Imputation	30.373
4	Regression Estimation	22.001
5	Expectation Maximization	15.543

3.3 Presentation of Results of Statistical Powers

Results of the statistical powers and non–centrality parameters for simulated data and two sets of real life data have been calculated and presented in tables 3.3.1 and 3.3.2.

Table 3.3.1: Statistical Power at various percentages of Missing Level

S/No	MV Estimation Method	5 %	10 %	15 %	20 %	25 %
1	Pairwise Deletion	0.902	0.836	0.863	0.905	0.729
2	Mean Substitution	0.933	0.917	0.959	0.981	0.951
3	Multiple Imputation	0.908	0.941	0.916	0.989	0.945
4	Regression Estimation	0.957	0.867	0.919	0.959	0.750
5	Expectation Maximization	0.931	0.927	0.918	0.992	0.981

Similarly, the results of the statistical power of the data of productivity improvement (PI) and Rehabilitation Therapy (RT) with the missing values and when the values have been estimated using various methods and presented in table 3.3.2

Table 3.3.2: Statistical Powers for the Data on PI and RT

S/No	MV Estimation Method	Productivity Improvement	Rehabilitation Therapy
1	Pairwise Deletion	0.998	0.999
2	Mean Substitution	1.000	1.000
3	Multiple Imputation	1.000	1.000
4	Regression Estimation	1.000	1.000
5	Expectation Maximization	1.000	1.000

3.4 Presentation of Results of Non–Centrality Parameters (NCP)

Non–Centrality Parameters for the hypothetical and real life data is presented in table 3.4.1

Table 3.4.1: Non–Centrality Parameters Results

S/No	Percentage missing	Non-centrality parameter
Hypothetical Data		
1	Complete data	18.082358

2	5 percent	18.134836
3	10 percent	18.193716
4	15 percent	18.260245
5	20 percent	18.336016
6	25 percent	18.423099
Productivity Improvement		
7	Complete data	16.950847
8	When data are missing	17.574779
Rehabilitation Therapy		
9	Complete data	17.316477
10	When data are missing	17.915433

4.0 Results and Conclusion

From our results, the following were observed:

1. Various methods exist for dealing with missing observations,
2. Mean Substitution (MS) method yielded the lowest variance (MSE) when the MVs were estimated.
3. Mean Substitution method yielded lowest MSE at 5, 10, 15 and 25 percent while at 20 percent Expectation Maximization had the lowest variance.
4. Pairwise Deletion method produced least statistical power when compared with the other Missing Value estimation methods.
5. Non–Centrality Parameters increased with increasing levels of missingness.
6. At 25 percent missing level, the statistical power reduced quite lower than for other percentage levels of missingness.

The following conclusions and recommendations can be made from this work:

1. When data are missing in one–way ANOVA, Expectation Maximization (EM) method should be used for the estimation. This is because, Mean Substitution (MS) method had the lowest variance (MSE), but because of its limitations which have already been pointed out by Cool (2000) and Little and Rubin (1987) it should not be used. Those limitations are:
   a. Over estimation of sample size,
   b. Negatively biased correlations,
   c. Underestimation of the variance and
   d. The distribution of the new values being an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean.

Expectation Maximization is therefore recommended because it was the method with the second least variances after MS.

2. Pairwise Deletion method should not be used since it reduces the statistical power of the results of ANOVA.

References

[1] Acock, A. C. (2005). Working with missing values. Journal of Marriage and Family, 67, 1012–1028.

[2] Arumuga N. S., (2015). A Comparative Study of Missing Value Imputation Methods on Time Series Data. International Journal of Technology Innovations and Research (IJTIR). Vol 14.

[3] Cool A. L., (2000). 'A Review of Methods for Dealing with Missing Data' Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 28, 2000.

[4] Eekhout, I., de Boer, M.R., Twisk, J.W.R., de Vet, H.C.W., and Heymans, M.W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology, 23(5), 729–732.

[5] Horton N. J. and Kleinman K. P., (2007) 'Much Ado about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models'. American Statistical Association

[6] Howell, D.C. (2008). The analysis of missing data. In Outhwaite, W. & Turner, S. Handbook of Social Science Methodology. London: Sage.

[7] Jain S., Jain K., and Chodhary N., (2016). 'A Survey Paper on missing Data in Data mining'. International Journal of Innovations in Engineering Research and Technology (IJIERT). 3(12).

[8] Little, R.J.A. & Rubin, D.R. (1987). Statistical analysis with missing data. New York: Wiley.

[9] Myers T. A., (2011) 'Goodbye, Listwise Deletion Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data' Journal of Communication Methods and Measures 5(4): 297–310.

[10] Myrtveit I., Stensrud E., and Olsson U. H., (2001). Analyzing Data sets with missing Data: An Empirical Evaluation of Imputation Methods and Likelihood–Based Methods. IEEE Transactions on Software Engineering, 27(11).

[11] Neter, J, Kutner, M. H, Nachtsheim, C. J, and Wasserman, W, (1996). Applied Statistical Models. MacGraw-Hill Companies 703-704.

[12] Peng C. J., Liou S., and Ehman L. H., (2003). Advances in Missing Data Methods and Implications for Educational Research.

[13] Rubin, D.B. (1976) Inference and missing data, Biometrika, 63: 581–92.

[14] Saunders J. A., Morrow-Howell N., Spitznagel E., Dori P., Proctor E. K., and Pescarin R., (2006). 'Imputing

Missing Data: A Comparison of Methods for Social Work Researchers'. Journal of Social Work Research 30(1).

[15] Schlomer G. L., Bauman S., and Card N. A., (2010). Best Practices for Missing Data Management in Counseling Psychology. Journal of Counseling Psychology, 57(1), pp. 1–10.

[16] Schmitt P., Mandel J., and Guedj M., (2015). A Comparison of Six Methods for Missing Data Imputation. Biometrics & Biostatistics.

[17] Sikka G., Takkar A. K., and Uddin M., (2010) Comparison of Imputation Techniques for Efficient Prediction of Software Fault Proneness in Classes. World Academy of Science, Engineering and Technology 38.

[18] Song Q., Shepherd M., and Cartwright M., (2005). 'A Short Note on Safest Default Missingness Mechanism Assumptions'. Journal of Empirical Software Engineering, 10, 235–243, 2005.

[19] Tanguma J., (2000). 'A Review of the Literature on Missing Data'. Paper presented at the Annual Meeting of the Mid–South Educational Research Association.

[20] Twala B., Cartwright M., and Shepherd M., (2006). Ensemble of Missing Data Techniques to Improve Software Prediction Accuracy. Journal of ICSE.

[21] Xu P., (2001). 'The Analysis of Missing Data in Public Use Survey Databases: A Survey of Statistical Methods'. A Thesis Submitted to the Faculty of the Graduate School of the University of Louisville in Partial Fulfillment of the Requirements for the Degree of Master of Science in Public Health Department of Bioinformatics and Biostatistics University of Louisville Louisville, Kentucky.