Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 477-492
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Study on the Missing Data Mechanisms and Imputation Methods

Abdullah Z. Alruhaymi, Charles J. Kim

Department of Electrical Engineering and Computer Science, Howard University, Washington DC, USA
Email: azmotairi@hotmail.com, ckim@howard.com

How to cite this paper: Alruhaymi, A.Z. and Kim, C.J. (2021) Study on the Missing Data Mechanisms and Imputation Methods. *Open Journal of Statistics*, **11**, 477-492.
https://doi.org/10.4236/ojs.2021.114030

Received: July 13, 2021
Accepted: August 8, 2021
Published: August 11, 2021

Abstract

The absence of some data values in any observed dataset has been a real hindrance to achieving valid results in statistical research. This paper aimed at the missing data widespread problem faced by analysts and statisticians in academia and professional environments. Some data-driven methods were studied to obtain accurate data. Projects that highly rely on data face this missing data problem. And since machine learning models are only as good as the data used to train them, the missing data problem has a real impact on the solutions developed for real-world problems. Therefore, in this dissertation, there is an attempt to solve this problem using different mechanisms. This is done by testing the effectiveness of both traditional and modern data imputation techniques by determining the loss of statistical power when these different approaches are used to tackle the missing data problem. At the end of this research dissertation, it should be easy to establish which methods are the best when handling the research problem. It is recommended that using Multivariate Imputation by Chained Equations (MICE) for MAR missingness is the best approach to dealing with missing data.

Keywords

Missing Data, Mechanisms, Imputation Techniques, Models

1. Introduction

Researchers have experimented with different methods that can tackle the missing data problem entirely for a long time. However, they have since explored different approaches and their effectiveness and determined that the old methods, such as univariate imputation and deletion methods, did not solve the problem and make the situation worse. This has led to invalid conclusions being arrived at in research because of incomplete data.

Researchers should, therefore, strive to abandon these custom methods in their projects and instead adopt the more modern practices of handling the missing data problem.

This article seeks to contribute to the search for the best missing data handling approaches by testing the different modern mechanisms of solving the problem since, when wielded correctly, data is an invaluable asset. However, that is only possible if the data utilized to develop a decision is accurate and conclusively. Data can help businesses gain leverage over their rivals. Bill Gates famously attributed Microsoft's dominance to the company's firm understanding of Bayesian belief networks, data-driven models used to aid decision-making. Beyond the worlds of academia and business, data has become an arsenal for governments and militaries. The cybersecurity attacks by groups purported to be supported by the Russian government led a United States Senator to state that the attacks were tantamount to a declaration of war [1]. This comes from the build-up of geopolitical tensions between world power vying for the latest technology to protect internal data while accumulating foreign data. The cybersecurity attacks on the Democratic National Convention (DNC) servers in the run-up to the 2016 US Presidential Elections have been argued as a leading cause in aiding the Republican win. These situations point to the significant importance of data and data integrity.

With complete data, experts can reverse engineer paths that led to incidents across industries such as aviation accidents or cybersecurity breaches, among others, to build the necessary infrastructure to prevent further incidents. However, the statistical power of the models weakens when there is missing data present, which can lead to complete misanalysis.

In cybersecurity, the main reason behind missing values in network security systems is network overloading and software failures. As a result, network databases often have missing values. For example, the KDDCUP'99 cybersecurity database may contain missing data in "tcpdump collectors", which are likely to become overloaded and drop packets under heavy traffic loads. Missing data is an obstacle when formulating an attack pattern detection model.

To understand how the missing data problem arises, it is important to note that statistical and machine learning methods have been developed over the years to analyse data sets for any incidences of missing data. These datasets are usually in rectangular form where its rows are usually composed of units of data known as cases and its columns consist of variables such as age, gender etc. which are typically measured for every case.

The entries in the rectangular dataset are in most cases composed of real numbers which either represent continuous variables such as income and weight or categorical variables such as gender and level of education.

If there are some entries in the dataset that are not observed, then the problem of missing data arises. Some of the main causes of this problem are, for instance, failure of a respondent to give answers to some of the questions asked in a sur-

vey or breakdown of equipment that are to measure some variables in our dataset such as temperature.

We aim to solve the following issues:

1) Impossible to find complete datasets; they are rarely discovered. Old univariate imputations and deletion did not solve missing data problems and even worsened results and gave the wrong outcome with invalid inferences and misled researchers and biased conclusions. Therefore, researchers should never be tempted to analyse the missing dataset. Also, it is imperative to distinguish missingness mechanisms to be able to select the best imputation method.

2) We should avoid missing data at the collection practice stage, and that is impossible, so we must use multivariate imputation by chained equation.

Analysing a dataset to make inferences is usually done when the dataset is in a rectangular format. This makes it easy to observe patterns of missing data and according to Schaffer and Graham, the identification of these patterns is a crucial step in classifying these patterns and eventually determining how to handle them.

Various research was conducted over the years, mainly after 2005. As researchers advance their knowledge of this domain, newer improvements were observed, reflecting extensive growth in Bayesian methods for simulating posterior distribution.

The work of Schafer & Graham [2] raised issues that remain unsolved to date, like MNAR mechanism conversion to MAR, analysis, and the use of auxiliary variables, and discuss dealing with other types of missingness. However, some of their work is still not yet in the mainstream.

White, Royston, and Wood [3] show how to impute categorical and quantitative variables, including skewed variables. They proposed and explained Multiple Imputation by Chained Equations as an immediate solution to missing data, which we will point to in the next chapter.

The authors Little and Rubin [4] represent approaches and the introduction of multivariate analysis with missing values. And lately Buuren [5] introduce Flexible Imputation of Missing Data and present MICE algorithm.

We mainly construct this Article for highlighting missing data mechanisms and briefly explain the multiple imputing processes; and in the next chapter, we extend our study to multiple imputation and the MICE algorithm.

We believe the added value of this Article to literature is the enforcement of some central concepts that are easily understood about missing data—contribute using modern techniques to avoid corruption of analysis of missing data in many fields, including Cybersecurity systems.

Researchers should try as much as possible to avoid analysing missing data by mainly adopting the best data collection and cleaning methods. It is, however, nearly impossible to altogether prevent this problem in many research projects. It is, therefore, crucial to find the possible mechanisms to decide to handle the missing data once identified.

2. Problem Statement

The missing data problem results if data values for one or even more data variables are not present from the observed dataset. A dataset is a group of observations which in most cases is rectangular and mainly denoted by $X=\left(X_1,\cdots,X_n\right)$ where $X_{i,i=1,2,\cdots,n}$ is the i^{th} column and $X^{j}{}_{,j=1,2,\cdots,m}$ is the j^{th} row. Suppose X_i is a randomly distributed variable which is defined by cumulative density function (CDF) $F_{k,k=1,2,\cdots,n}$, and every X_i has a different probability density function (PDF). Then the cumulative density function of $X_{i,i=1,2,\cdots,n}$ should be defined as; $F\left(X_i\right)=Pr\left(X_j\le X_i\right)$. If the observed dataset X has missing values denoted by Y_p^q, then this is missing value of the q^{th} and p^{th} columns. Also, $Y_{(m_1,n),n_1<n}$, is a group of rows containing missing data values, $Y_{(m_1,n)}=\left\{Y_p^q\right\}$ and hence $Y_{(m_1,n)}\subset X$.

To clearly indicate the amount of the missing values and their positions, let R be the indicator of the missing value whose elements have the values one and zero. When $R=1$, this indicates that the data within the dataset is known while when $R=0$, this indicates that the value is missing. Also let Y_m be the number of the missing values in a certain row j and Y_0 be the number of known values in the same row, then $Y_m=\sum_{R=0}X^R$ and $Y_0=\sum_{R=1}X^R$. Therefore, if the missing element has a null value, then the observed dataset is complete, or else it has some missing values.

3. Classification of Missing Data

Rubin [6], who is a pioneer in this field, asserted that missing data is based on why they are missing from the beginning? Investigating reasons of absence, and ought to be classified as one of the following three mutually exclusive categories:

- Missing at random
- Missing completely at random
- Missing not at random

It is not always clear why data is missing; however, some inferences can be made by detecting the pattern of the missing data. In addition, guidelines provided by Rubin and later Little [4] have provided researchers with enough information to develop model's derivatives of those first created by the originators.

According to Rubin [7], where a pattern in the missing data is unobservable, there is no apparent correlation of the values missing to the dataset itself, and the data is classified as MCAR. Little [8] developed what would be known as Little's MCAR test that investigates the null hypothesis that the observed missing data is of the MCAR category by calculating the p-value of the missing data and then comparing it to significance level of the test. And if the obtained p-value is less than the significance level then it is concluded that the missing data is not of the category MCAR.

MAR holds because, for this mechanism, data depends on the values that are observed and are independent of unobserved ones; thus, performing a test of MAR against MNAR is impossible because it requires unavailable data, as stated

by Buuren [9]. At the same time, the MCAR method assumes that the dataset does not depend on the observed and unobserved values and can be tested by Little's test for the observed values. And if the missing data assumption is dependent only on the unobserved data, then the mechanism of the missing data pattern is MNAR (Table 1).

3.1. MCAR Data

Data that is categorized as (MCAR) does not depend on the observed and unobserved data [10]. Particularly, there exist no systematic differences between respondents with incomplete data and those having complete data. For instance, some data records may have incomplete (missing) information because of a mishap at a laboratory or errors during data entry. Although the data is missing and the sample size and power to see significant findings are reduced, bias is not introduced. In essence, with MCAR data, one can assume that the missing data, if somehow found and added to the dataset, would not change the conclusions reached during analysis. Thus, it can be assumed if data is indeed MCAR, there is no bias introduced to the data that remains. One can then think that the data that does remain is a simple random sample of the full dataset and representative of the entire data set.

MCAR is usually considered as a strong yet repeatedly non-realistic assumption. MCAR data is also the easiest to account for during data preparation and analysis. One option of adjusting for MCAR data is to remove all records with missing data from the dataset (listwise deletion). This is because MCAR data is similar in scope to the remaining data and removing the records with missing information will not bias the results obtained from the data that remains.

The percentage of data that is missing is also a factor in whether the missingness can be considered MCAR. However, experts disagree on the portion of missingness in which MCAR can be assumed. Schaffer [11] suggested that if 5% or less of the data is not in the whole dataset, this missing data can be considered MCAR. Other researchers have suggested that MCAR can be assumed for missingness in a dataset when the amount of missing data is 10% or less [12], and even 20% or less.

The percentage of the missingness should not be the only determinant of MCAR data. Of more importance is the pattern of missingness and the size of the dataset that remains when the missing records are removed. As mentioned earlier, the power of a study will be reduced when records are removed. And reduced power may result in Type I error (not detecting statistically significant findings when the significance is truly present.

Table 1. Missing data mechanisms explained.

Assumption	*Observed*	*Unobserved*
MAR	Dependent	Non-dependent
MCAR	Non-dependent	Non-dependent
MNAR	Non-dependent	Dependent

3.2. MAR Data

When data is missing at random (MAR), its absence is analytically more likely to be related to the observed than the unobserved data [13] e.g., the likelihood of detecting anomaly intrusions might depend also on a dataset's prevention system.

3.3. MNAR Data

When data is missing not at random (MNAR), its absence in this case is analytically more likely to be related to the unobserved data, and this means its missingness is related to factors that cannot be accounted for or measured by the researcher [14].

Listwise deletion of records from a data set comprising of the MNAR data has the possibility of resulting in biasness; if, however, the complete case analysis is already biased it means that the sources of missing data are themselves not accounted for suggesting that the missingness cannot be measured in analysis and any results obtained will therefore likely be biased. When we understand the missing data, we can deal with it and know how to treat it.

4. Missingness Explanation

Dealing with missingness is essential, and indeed one needs to understand the mechanisms and patterns of missing data. Attributes that did not contain missing data are called complete data, while features that have missing values are called incomplete data. For example, referring to the 2.2 of the Dissertation, Cyber database selected for this research, is the KDDsubset reduced to cleansing data, a complete data consisting of 145585 connections and 39 features. The label column was excluded from making the data incomplete at a different proportion of 10%, 20%, and 30% with two missingness mechanisms: MCAR and MAR.

Then to feed the imputed missing data, the label returned to the dataset to test by a Machine Learning algorithm to measure the accuracy for each classifier; we did R code to analyze data and make it incomplete by MCAR. This mechanism means the data in the specific column or other columns have no relationship because the missingness is unsystematic. Each cell of the dataset has the same chance to be selected equally likely to the sample, so we randomly sample from our data set as the population and have sample missing datasets with the mentioned proportions.

For MAR missingness, it is a bit more complex because the missingness depends on the observed value, and for this reason, we wrote an R code that was to make MAR is conditional MCAR, whereas the probability of cells in MCAR missingness is the same, and samples are randomly selected. But the probability in MAR is conditioned and hence called CMAR (Conditional Missing at Random) not to be confused with the two abbreviated terms. Probability is given some conditions and equal to MCAR. Quantiles are used as the condition for

this assumption. Finally, we imputed the missing data using the MICE algorithm. Then we fed it to the Machine Learning algorithm to measure accuracy, these were done in different chapters; for MNAR the most problematic type is left for further investigation research (Figure 1).

4.1. Detection of the Type of Missingness

Statistical models for handling missing data rely on assumptions e.g., data is MCAR. These assumptions that govern the missingness of the data must be inspected prior to imputation. Multiple Imputation by Chained Equations (MICE) methods are heavily reliant on the assumption of missing values being MCAR or MAR. To verify, it is necessary to inspect and test the data. It is quite common for professional surveyors to, follow-up on a paper survey with phone calls to the non-respondent's group and enquire on some important survey items. This would allow for comparisons to be made between respondents and non-respondents. Chasing up to handle missing data is not always possible; the task may be economically unfeasible and missing data may come from anonymous responses with no way of contacting the survey respondents. For these reasons, uncertainty in the knowledge of why data is missing will always exist. Statisticians developed models to help identify the cause of the missingness, notably Little's test [15] for MCAR detection, and the likelihood-ratio test for MAR.

4.2. Diagnosing the Missingness Mechanisms

When data is classified as MNAR, it is deemed to be not easy to disregard, as the missing data method itself ought to be used to influence the type of models that may be used on that data. Where data is classified MCAR or MAR, no information is required about the missing data, thereby considered ignorable. It is necessary to understand the mechanism type to get valid statistical results. The distinction may be necessary in determining how missing data is treated, when proportion of missing data is known. For example, if the absent data is categorized

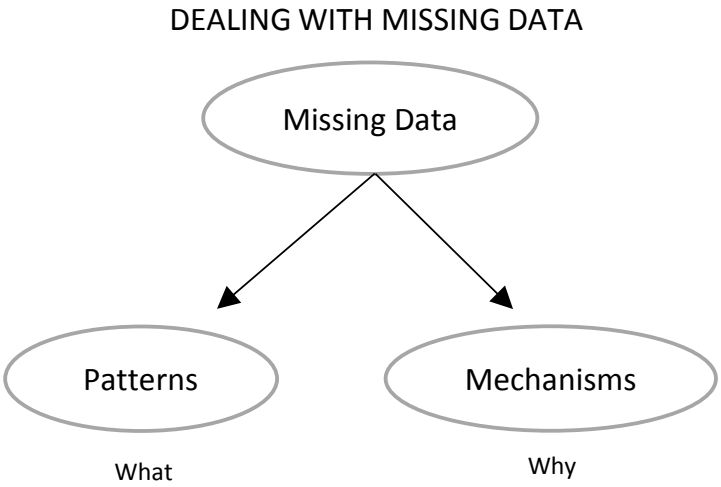

Figure 1. What and why missing data.

as MCAR then listwise deletion may be considered appropriate should it be less than 10% of the data to mitigate against bias induction in models.

1) MCAR VS. MAR

Little's test is frequently used for MCAR; however, as with all missingness tests it is not considered absolute. That is, a problem of misclassification of the mechanism cannot always be spotted. Alternatively, another method of attempting to identify the classification is by creating dummy variables, coding missing data as 1 and observed as 0. This could allow for t-tests to be performed, followed by chi-squared tests of the data.

2) MNAR VS. MAR

The mechanisms MNAR and MAR have considerable overlaps, making it easy to easily misclassify one mechanism for the other. In the context of a survey/questionnaire the preferable option would be to follow up with a survey and or questionnaire, however, as previously mentioned this method is likely to face obstacles of cost and or feasibility.

3) MCAR VS. MNAR

MCAR means there isn't any relationship between the missing data and the other observed values. This means there isn't anything special that makes some of the observed data more likely to be absent than the different data values, and it is random missingness. MNAR, on the other hand, means there is a relationship of the missingness of some of the datasets and requires future work on the Bayesian theory analysis and new approaches to correlation links to figure out reasons beyond missingness and how missingness depends on itself.

5. Literature Review

A few publications such as Little and Rubin [16] and Schafer [17] give an incredibly detailed and sophisticated theoretical outlook and analysis of the missing data. Rubin [18] and Schafer [19] avails a complete discussion of the theoretical conjecture of multiple imputation, and some examples of its use.

Stef van Buuren, [20] focused on the flexible nature for the many processing data types present in a lot of real applications in the multiple imputation framework. The technique for creating various estimates of missing data values varies but multivariate imputation by chained equations (MICE), a popular method creates these estimates using: Predictive mean matching, logistic regression, Bayesian linear regression, and many others (Buuren and Groothuis—Oudshoorn) [21]. Multivariate imputation by chained equations from Stef Buuren has arisen in the data analysis as one robust major technique of handling missing data. Generating multiple imputations, instead of single imputations, account for the collection of uncertainty in data in the imputations. In addition, the chained equations method is so flexible and able to handle variables of different types, including continuous data. This method assumes that the missing data is a MAR type (missing at random). And this means that the likelihood of a being missing from a dataset is solely dependent on the observed values. In other words, the

missingness that remains after the control of all the available data is purely random.

6. Tests of Missing Data

Before we delve into missing data definitions, it is crucial to establish a ruler for the missing data method because of the unseen variety of differences between the three assumptions. Maybe MCAR is the most straightforward assumption, meaning that it exists when the missing data values are distributed randomly through observations, and confirmation can be done by partitioning the dataset. Sample into two sets: one set comprising the missing data values and the other containing the non-missing data values. Then use a t-test of mean difference to determine whether there is any difference in the samples among the two datasets. If the data is MCAR, we may use pairwise or list-wise deletion for the cases of the missing value depending on the proportion of the missing itself. If, on the other hand, the data is not MCAR, then we conduct multiple imputation by chained equations (MICE) algorithm to impute the missing data. In MAR, the missingness is not random as in MCAR, and missing data is distributed throughout the observations but distributed in one or more sub-samples. To distinguish between the three missing data mechanisms, we introduce some traditional methods of identifying missing data mechanisms and tests that may assist in defining the nature of the mechanism type. However, missing data mechanisms are not controlled by the investigators and need future exploration and deep thinking.

- P-value is the probability of getting an outcome that is at least as extreme as the observed one. When p-value < 0.05 then there is significant statistical difference, and when p-value > 0.05 there is no significant statistical difference.
- T-test is used as a hypothesis testing to allow testing of an assumption applicable to a population.
- Likelihood-ratio test is a hypothesis that is helpful in choosing a model that is better between two models. A test is proposed by Diggle [22] for MAR assumption but as for the p-values are distributed under the null hypothesis, Diggle refers to use Kolmogorov's test [23] to decide the resulting p-value, p_j behaves like a random sample from a uniform distribution.

We can calculate (λ) for the dataset and compare: $-2\log(\lambda)$ to χ^2 value corresponding to a desired statistical significance.

- Chi-Square test compares two categorical variables to see if they are related, it is used for two purposes; goodness of fit to ensure that the sample dataset matches the population, and for independence when comparing variables. Usually, we use contingency tables to determine if one variable has a particular effect on the other by calculating the observed and expected value, then perform the chi square test to see if the observed values fit expected values well. Same steps that we use for hypothesis, are also used for chi square test

and these are: stating null and alternative hypothesis, choosing the level of significance (α) as it is just the area of the tail as in the shown diagram, finding the critical values, finding test statistic, and then drawing a conclusion about the null hypothesis. Rejecting it means we currently believe that the alternative hypothesis is true. To look for the critical value in the chi square table we must know the degrees of freedom, calculated by (rows − 1) * (columens − 1) or (n − 1) and look for the result that intercepts with column $\chi^2 = 0.05$. Then find test statistic which is the Chi square we use the formula $\left(\chi^2\right) = \text{sum}\frac{(O-E)^2}{E}$, and if it is less than the critical value, we draw a conclusion that we cannot reject the null hypothesis, so we accept it (**Figure 2**).

From the table below we observe that Chi square increases as the (df) degree of freedom decreases (**Table 2**).

- Dixon's MCAR test compares the means of complete and incomplete cases, and if the t-statistic is insignificant then the data is MCAR. If, however, the t statistic is significant the data is not MCAR.
- Little's test for MCAR data is a test that is used widely to determine if data can be assumed to be MCAR (Little, 1988). Little's MCAR test is the most significant test when dealing with missing cases being MCAR. If its p-value is statistically insignificant, then the data is assumed to be MCAR, and missingness is then assumed not to be important in the analysis. Using listwise deletion method on observations with missing data values is then appropriate approach to deal with them, or we can use the most advanced relabel imputation method, multiple imputation by chained equations if a more complete dataset is needed to increase the sample size and achieve better statistical power.
- Hawkin's test for MCAR, is a test of multivariate normality as well as a test of homogeneity of covariances.
- Wilks's test shows the amount of variance accounted for in the response variable by the explanatory variables.

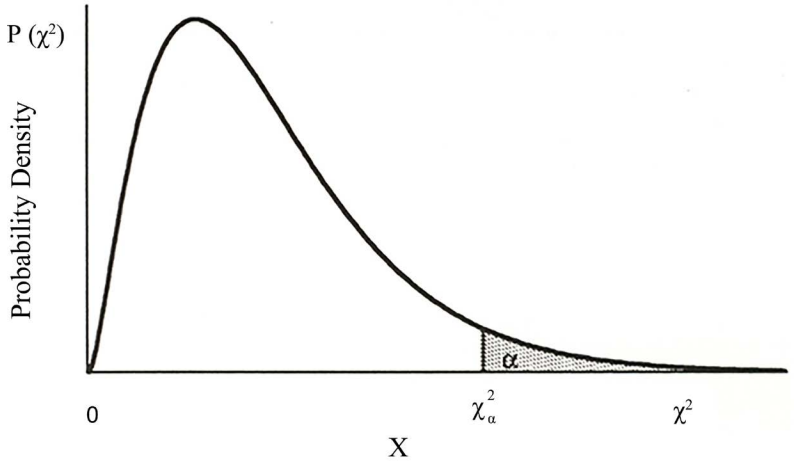

Figure 2. A curve shown the rejection region (tail) [24].

Table 2. Chi square intercept with degree of freedom [25].

df	X^2 0.995	X^2 0.990	X^2 0.973	X^2 0.950	X^2 0.900	X^2 0.100	X^2 0.050
1	0.00	0.000	0.001	0.004	0.016	2.706	3.814
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070

- For MAR there are other tests like (Diggle's test, Kolmogorov MAR test, and Fisher test).
- For MNAR there is Fairclough approach.

7. Probability of Missingness Definitions

No causal clarity for the missing data when considering mechanisms, but they offer mathematical relationships between the data observed and the missingness represented in the following equations for each mechanism's probability:

Let: Y_o = data observed, Y_m = data missing, R = response indicator 0-1 matrix, X = covariate values, and $Y = Y_o + Y_m$.

The data is said to be MAR and its defined as:

$$Pr\left(R \mid Y_m, Y_o, X\right) = Pr\left(R \mid Y_o, X\right) \tag{1}$$

For Equation (1), we use the likelihood-ratio test (λ) and from the dataset we calculate the mean and standard deviation to see if the mean (μ) is equal to a certain estimated value (μ_0). The null hypothesis $H_0 \neq H_1$ where $H_0 = \mu$ then the likelihood function after calculation would be [26] $\lambda = \left(1 + t^2/n - 1\right)^{-n/2}$. Where t is the statistic with $n-1$ degrees of freedom.

If our interest is testing if Y is MCAR or not, according to the Little's test of MACR we should indicate that p-value is less than 0.05 (<0.05) meaning that there is weak evidence against the null hypothesis, so we fail to reject the null hypothesis (H_0) and in this case the null hypothesis is that the missing data is MCAR [8]. Because there exist no patterns that exist in the missing data assumption. Proving on the other hand, the existence of the MAR assumption is very difficult, but we could try if the data is related between the two assumptions. The best solution for this problem is to look about the data and understand the definition of MCAR if it is plausible for the data to be MCAR and this will prove so.

Denote that the observed and missing entries of Y as Y_o and Y_m, respectively. The assumption for MCAR is weaker than that of MAR and is defined as follow:

$$Pr\left(R \mid Y_m, Y_o, X\right) = Pr\left(R\right) \tag{2}$$

If the missing data approach is not satisfying Equation (1) MAR type, then the assumption here is MNAR.

$$Pr\left(R \mid Y_m, Y_o, X\right) = Pr\left(R \mid Y_m, X\right) \tag{3}$$

8. Missing Data Imputation

Missing data techniques can be divided into simple traditional techniques and modern techniques; and of course, the current methods are always preferable, but there are some cases where the old techniques may be applicable. Also, we need to understand why the simple techniques do not adequately work.

8.1. Traditional Methods

Some older methods to handle missing data include deletion and single imputation approaches [27]. A complete-case analysis, which is list-wise deletion, discards missing values entirely from the dataset. Pair-wise deletion removes only incomplete cases as a remedy to the data loss of list-wise deletion (**Table 3**).

This method helps with small missing data in the dataset and may not cause bias to the analysis. However, it will deprive important information in the research, though it used to be the most common method in quantitative research to treat missing data. Imputing data with mean/ median is another quick fix for missing data, but it causes bias in the analysis since it decreases the variance of data points.

Table 3. Examples of list wise deletion and pair wise deletion [28].

List wise deletion		
Gender	Manpower	Sales
M	25	243
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
.	26	259
M	32	297

Pair wise deletion		
Gender	Manpower	Sales
M	25	243
	.	
M	33	332
	.	272
		.
M	29	326
.		
M	32	297

As for imputation with linear regression, few predictors of the variables with missing values are recognized with the help of a correlation matrix. The best predictors are used as predictor variables and the variable with missing data as the dependent variable in the regression equation, which is then used in the prediction of missing values. Imputation with stochastic regression improves the linear regression-based imputation, which adds random noise terms on a regression line to restore lost variability to the data.

Other traditional techniques include general techniques, scale-item techniques, and time-series techniques.

Enders summarized the traditional techniques for missingness mechanisms that we consider like a road map for handling missing data as in Table 4.

All the above-mentioned methods, however, cause bias and do not satisfactorily solve the problem of missing data.

8.2. Modern Imputation Techniques

The most challenging task is dealing with missing values because the exact nature of the missing data is unknown. The "State of the art" methods that dealt satisfactorily with handling missing data are multiple imputations and maximum likelihood. In this Dissertation, we will concentrate on multiple imputations and how it works.

1) Multiple imputations [30] is considered as the best approach for dealing with the missing data problem as it creates several copies of the dataset, each containing different imputed values.

2) Maximum-likelihood imputation [31], which works around a variance-covariance matrix for the variables based on all the available data points.

Multiple imputations produce unbiased estimates with MCAR and MAR data. Handling missing data using multiple imputation methods has various benefits to it. The major benefit is that these methods reduce the biasness in the dataset when doing analysis of the data. They also increase the accuracy of the data and thus advance on the validity of an experiment being undertaken using the data. Again, there is an increase in the precision of the data as the imputation methods bring closer the data values within a dataset. Lastly, imputation methods lead to a more robust statistical analysis since they make a dataset more resistant to outliers.

Other methods include the expectation-maximization algorithm and the Bayesian simulation method [32].

A recent imputation approach known as MICE is widely accepted in the data

Table 4. Summary of traditional ways of treating missing data [29].

MCAR	MAR	MNAR
• Listwise deletion if sample size is very small	• Listwise deletion if amount is not important, and small • Use stochastic regression imputation	• None of the above applicable • Selection models • MAR techniques

imputation field [33]. Therefore, in this dissertation, the MICE technique, along with multiple imputations, has been explored and adopted. Details regarding multiple imputations and MICE and its application to the analysis are described in detail in a subsequent chapter.

9. Conclusion

Missing data is always a limiting factor when undertaking any test or an experimental project, regardless of whether the missing data is MAR, MNAR, or MCAR. There will always be a loss of statistical power in all these scenarios that can lead up to a type I error. And this may result in the making of inaccurate inferences about a population. Therefore, researchers should try and prevent missing data patterns in the dataset used in their research if possible. This can only be achieved through a careful approach to data collection and preparation, as stated earlier in the chapter. On the other hand, if these missing data patterns are impossible to avoid, they are in most cases. Then researchers should adopt appropriate techniques for handling the missing data and, more preferably, modern methods such as the multivariate imputation by chained equations (MICE) because traditional techniques involve excluding cases that have missing values in the dataset. This old method is inappropriate since research results often strive to make inferences about an entire population and not just a portion in a dataset.

Acknowledgements

The authors would like to thank reviewers for their helpful notes and suggestions to this article. Thanks, to extend to Scientific Research/ the Open Journal of Statistics for their valuable reviews and publishing.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Brangetto, P. and Veenendaal, M.A. (2016) Influence Cyber Operations: The Use of Cyberattacks in Support of Influence Operations. 2016 8*th International Conference on Cyber Conflict* (*CyCon*), Tallinn, 31 May-3 June 2016, 113-126. https://doi.org/10.1109/CYCON.2016.7529430

[2] Schafer, J.L. (2003) Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Neerlandica*, **57**, 19-35. https://doi.org/10.1111/1467-9574.00218

[3] White, I.R., Royston, P. and Wood, A.M. (2011) Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, **30**, 377-399. https://doi.org/10.1002/sim.4067

[4] Little, R.J. and Rubin, D.B. (1989) The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research*, **18**, 292-326.

https://doi.org/10.1177%2F0049124189018002004

[5] Van Buuren, S. (2018) Flexible Imputation of Missing Data. CRC Press, Boca Raton.

[6] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, **63**, 581-592. https://doi.org/10.1093/biomet/63.3.581

[7] Rubin, D.B. (1978) Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Vol. 1, American Statistical Association, Alexandria, 20-34.

[8] Little, R.J. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, **83**, 1198-1202. https://doi.org/10.1080/01621459.1988.10478722

[9] Doove, L.L., Van Buuren, S. and Dusseldorp, E. (2014) Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects. *Computational Statistics & Data Analysis*, **72**, 92-104. https://doi.org/10.1016/j.csda.2013.10.025

[10] Chen, H.Y. and Little, R. (1999) A Test of Missing Completely at Random for Generalised Estimating Equations with Missing Data. *Biometrika*, **86**, 1-13. https://doi.org/10.1093/biomet/86.1.1

[11] Schafer, J.L. and Olsen, M.K. (1998) Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, **33**, 545-571. https://doi.org/10.1207/s15327906mbr3304_5

[12] Graham, J.W. and Hofer, S.M. (2000) Multiple Imputation in Multivariate Research. In: Little, T.D., Schnabel, K.U. and Baumert, J., Eds., *Modeling Longitudinal and Multilevel Data*, Psychology Press, New York, 189-204. https://doi.org/10.4324/9781410601940-15

[13] Heitjan, D.F. and Basu, S. (1996) Distinguishing "Missing at Random" and "Missing Completely at Random". *The American Statistician*, **50**, 207-213. https://doi.org/10.1080/00031305.1996.10474381

[14] McPherson, S., Barbosa-Leiker, C., Mamey, M.R., McDonell, M., Enders, C.K. and Roll, J. (2015) A 'Missing Not at Random' (MNAR) and 'Missing at Random' (MAR) Growth Model Comparison with a Buprenorphine/Naloxone Clinical Trial. *Addiction*, **110**, 51-58. https://doi.org/10.1111/add.12714

[15] Little, R.J. and Smith, P.J. (1987) Editing and Imputation for Quantitative Survey Data. *Journal of the American Statistical Association*, **82**, 58-68. https://doi.org/10.1080/01621459.1987.10478391

[16] Little, R.J. and Rubin, D.B. (2019) Statistical Analysis with Missing Data. Vol. 793, John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119482260

[17] Graham, J.W., Hofer, S.M., Donaldson, S.I., MacKinnon, D.P. and Schafer, J.L. (1997) Analysis with Missing Data in Prevention Research. https://content.apa.org/doi/10.1037/10222-010

[18] Rubin, D.B. (2003) Discussion on Multiple Imputation. *International Statistical Review*, **71**, 619-625. https://doi.org/10.1111/j.1751-5823.2003.tb00216.x

[19] Schafer, J.L. and Graham, J.W. (2002) Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, 147-177. https://doi.apa.org/doi/10.1037/1082-989X.7.2.147

[20] Van Buuren, S. (2011) Multiple Imputation of Multilevel Data. Routledge, 181-204. https://doi.org/10.1201/b11826

[21] Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., Jolani, S., *et al.* (2015) Package 'Mice'.

[22] Diggle, P.J. (1979) On Parameter Estimation and Goodness-of-Fit Testing for Spatial Point Patterns. *Biometrics*, **35**, 87-101. https://doi.org/10.2307/2529938

[23] Barr, D.R. and Davidson, T. (1973) A Kolmogorov-Smirnov Test for Censored Samples. *Technometrics*, **15**, 739-757. https://doi.org/10.1080/00401706.1973.10489108

[24] Saylordot Organisation (2019) 11.2 Chi-Square One Sample Test of Goodness of Fit. https://saylordotorg.github.io/text_introductory-statistics/s15-02-chi-square-one-sample-goodness.html

[25] Tallarida, R.J. and Murray, R.B. (1987) Chi-Square Test. In: *Manual of Pharmacologic Calculations*. Springer, New York, 140-142. https://doi.org/10.1007/978-1-4612-4974-0_43

[26] Kent, J.T. (1982) Robust Properties of Likelihood Ratio Tests. *Biometrika*, **69**, 19-27. https://doi.org/10.1093/biomet/69.1.19

[27] Scheffer, J.A. (2000) An Analysis of the Missing Data Methodology for Different Types of Data: A Thesis Presented in Partial Fulfilment of the Requirements for the Degree of Master of Applied Statistics at Massey University. Doctoral Dissertation, Massey University, Palmerston North.

[28] StackOverflow (2017) Machine Learning with Incomplete Data. https://stackoverflow.com/questions/39386936/machine-learning-with-incomplete-data

[29] Enders, C.K. (2010) Applied Missing Data Analysis. Guilford Press, New York.

[30] Jakobsen, J.C., Gluud, C., Wetterslev, J. and Winkel, P. (2017) When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomised Clinical Trials—A Practical Guide with Flowcharts. *BMC Medical Research Methodology*, **17**, Article No. 162. https://doi.org/10.1186/s12874-017-0442-1

[31] Graham, J.W., Hofer, S.M. and MacKinnon, D.P. (1996) Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures. *Multivariate Behavioral Research*, **31**, 197-218. https://doi.org/10.1207/s15327906mbr3102_3

[32] Andradóttir, S. and Bier, V.M. (2000) Applying Bayesian Ideas in Simulation. *Simulation Practice and Theory*, **8**, 253-280. https://doi.org/10.1016/S0928-4869(00)00025-2

[33] Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J. (2011) Multiple Imputation by Chained Equations: What Is It and How Does It Work? *International Journal of Methods in Psychiatric Research*, **20**, 40-49. https://doi.org/10.1002/mpr.329