Sirte University Scientific Journal Vol. 14, No. 2 (2024) 57-64

Sirte University Scientific Journal (SUSJ)

Journal home page: http://journal.su.edu.ly/index.php/SUSJ/index

DOI: 10.37375/susj.v14i2.3090

A Study on Evaluation of Mean and Median Imputation Methods and Their Impact on Statistical Analysis of Missing Data

Fatma M. Kikhia

Department of Statistics, Faculty of Science, University of Benghazi, Libya.

DOI: 10.37375/susj.v14i2.3090

ARTICLE INFO:

Received 22 October 2024.
Accepted 30 November 2024.
Available online 24 December 2024.

Keywords: Missing Data Imputation, Mean Imputation ,Median Imputation, Handling Missing Values.

A B S T R A C T

In this study, we are going to evaluate the effect of employing diverse missing value imputation strategies on real datasets while conducting statistical analysis. It will concentrate on Missing Completely at Random (MCAR) data, which provides an opportunity for a thorough assessment of imputation methods. Specific methods examined were mean and median imputation plus other conventional statistical ways of treating missing data. As a result, the research underlines that adequate data management strategies are key to preserving both the credibility and accuracy of scientific analyses. This study demonstrates how Excel can be used as the primary analytical tool to give applied researchers from different areas of specialization practical guidance on method choice when faced with missing data. In the end, these results demonstrate how carefulencial efforts are essential in this field.

Introduction

Reliable statistical analyses require accurate and complete data, as incomplete information can lead to biased estimates and incorrect conclusions. This is a challenge that researchers across disciplines—whether in medicine, social sciences, or economics—often encounter. As emphasized by Schafer and Graham (2002), missing data can compromise the validity of statistical analyses, particularly if not handled appropriately. Such challenges highlight the importance of following established procedures for managing missing data to ensure the reliability of research outcomes.

Therefore, in this paper, we set an objective to assess the effect of various imputation methods applied on real and complete datasets. More specifically, we decided to create missing data using the Missing Completely at Random (MCAR) mechanism to make it possible for us evaluate how well and efficient different methods of imputation can take care of such cases. Although not all methods will be investigated in this study, common techniques like replacing missing values using a mean or median is the most important step. Previous studies have highlighted the impact of missing data on statistical analysis, demonstrating that the proportion of missing values can significantly affect the results, particularly in relation to sample size. This study underscores the importance of understanding these dynamics to implement effective imputation strategies in diverse research contexts.

All analysis will be performed using Microsoft Excel, enabling non-programming researchers to apply these techniques in practice. This study is written to provide incremental results for choosing specific missing value techniques in different data environments addressing a gap identified by Graham (2009), who highlighted the need for accessible tools in missing data analysis. — and hence the other side of paper may not be so suitable all the time.

The Purpose of the Study

1. Evaluate the effectiveness of imputing missing data using mean-median replacement:

The purpose of this study is to examine the effectiveness of a method of estimating missing values, especially replacing missing values with mean and mean. Although research accepts other methods such as regression technique and multiple imputation, but will focus primarily on the aforementioned method.

2. Examine the effect of missing data on statistical analysis results:

The study seeks to investigate how the proportion of missing values and the chosen method of comparison affect the results of statistical analysis, particularly the effect on key statistics such as the mean and standard deviation.

3. Make data analytics tools easily accessible:

This study will use Excel as a tool for data analysis, making statistical analysis methods accessible to researchers who may not have programming backgrounds.

4. To increase understanding of the impact of missing data on statistical analysis:

The study aims to improve both academic and professional understanding of the impact of missing standards in statistical analysis, thereby helping to advance data quality in future research efforts.

Literature Review Introduction

Missing information in data sets poses significant challenges for researchers in fields such as health care, social sciences, and economics. Missing variances can lead to biased estimates and compromise the integrity of the statistical analysis. According to Little and Rubin (2002), missing data can lead to "biased parameter estimates and reduced statistical power" (Little & Rubin, 2002).

Missing data are classified into three categories: completely missing at random (MCAR), missing at random (MAR), and nonrandom (MNAR). Each stage presents unique challenges that require different analytical approaches. Traditional methods such as mean and median imputation have been used extensively but often infrequently. For example, mean imputation can distort data distributions and reduce variability, as highlighted by Van Buren and Groothuis-Odschoorn (2011), who state that "it can lead to underestimation of standard errors" (Van Buren & Groothuis). -Odschoorn, 2011) In contrast, multiple imputation provides a more robust alternative by generating multiple data sets reflecting the uncertainty of missing values.

In this literature review we focus on a few different methods to deal with missing data, especially methods of estimating means and medians and other methods not currently covered in this review In addition to findings in previous research together, this study attempts to clarify the current situation, finding gaps in literature. Finally, this discussion introduces current theory using real international datasets that look at the impact of implied mean imputation on statistical analysis results.

Definition of Missing Data

Missing data means there are no observations in the dataset, posing a significant challenge to statistical analysis. as Schafer (1997) highlights. Missing data can significantly affect the validity of research findings. Missing information is generally classified into three main types:

1. Missing completely at random (MCAR):

• This trend occurs when missing observations are not associated with any measured value. Basically, the missing data is completely random. For example, if study participants chose not to answer a particular question for which there was no discernible reason, the statements could be classified as MCAR.

2. Missing at Random (MAR):

• MAR occurs when the missing data is associated with other observed values in the data set. For example, if individuals with a particular health condition are less likely to answer a survey question, this indicates that missing data are dependent on other available data.

3. Not missing at random (MNAR):

• MNAR refers to cases where missing observations are associated only with missing values. For example, patients experiencing more severe symptoms are less likely to participate in assessments, and information about the severity of their condition is lacking.

The occurrence of missing data can result from a variety of factors, such as errors in data collection, reluctance of participants to disclose personal data, or technical problems with instruments the data collection process.

Strategies for Dealing with Missing Data

Dealing with missing data is an important challenge in statistical research, leading researchers to take various approaches to address this issue, Effective management of missing data is crucial, as emphasized by Van

Buuren (2018) and Enders (2010). These strategies can be divided into several types:

1. Deletion methods:

• Listwise deletion: This method removes any rows with missing values from the search. Although simple, it can lead to significant data loss and bias if the missing data is not random.

• Pairwise deletion: This method removes missing values for specific variables during analysis, allowing researchers to save more data than listwise deletion but creating inconsistencies that make the results interpretable is strong.

2. Mean/median imputation:

• Researchers often replace the mean or median of observed values with missing values for those variables. Although this method is simple and easy to implement, it can reduce the variability in the data set and distort the relationships between variables, especially when the absence is not random Frequency is used with mean imputation, where the mean of available values is calculated to compensate for missing data. Imposing the median is another method that can be more robust with outliers.

3. Regression imputation:

• This method is based on observed values using regression models to predict missing values. Although the regression approximation may provide more accurate estimates than the average, it may reduce the uncertainty associated with missing data.

4. Multiple allegations:

• Multiple imputation is a more advanced method that generates multiple data sets with different imputed values based on observed data. Each data set is analyzed separately, and the results are combined to account for the uncertainty of missing values. This method yields valid statistical parameters and is highly recommended in research.

5. Maximum Likelihood Estimation (MLE):

• MLE estimates the parameters of a statistical model to increase the likelihood of identifying observations while considering missing data. This method is versatile and can be adapted to different circumstances.

6. Machine Learning Methods:

• Recent advances in machine learning have led researchers to use techniques such as k-nearest neighbor (KNN) and random forests to predict and estimate missing values based on patterns in data. These methods can often outperform conventional allergy methods.

7. Chained Equations:

• This method models missing data through multiple regression models, improving the accuracy of imputed values. Chained equations provide a simple and efficient framework for dealing with missing data.

8. Group based imputation:

• In cluster analysis, missing values are calculated based on the characteristics of the same data point or cluster. This method can work well for datasets with natural clusters, although it may degrade in performance if the clusters are poorly defined or large amounts of data are missing

Choosing the appropriate comparison method depends on the amount of missing data, the nature of the data set, and the purpose of the analysis Although simpler methods such as averaging or median imputation may be effective for small data a there is a small amount of missing value though sophisticated methods such as repeated imputation.

Previous Studies

Several researchers have examined the issue of missing data and proposed various strategies to deal with it. For example, Graham (2009) highlighted the potential for bias caused by even small amounts of missing data, particularly in fields such as psychology, if not properly addressed. His work emphasized the importance of comprehensive methods, such as multiple measures, to reduce bias and improve the accuracy of the analysis.

Schafer and Graham (2002) also provided a detailed analysis of missing data techniques including deletion, mean imputation, and regression techniques, examining their strengths and weaknesses Their findings suggest that although methods that do not difficult are easy to implement though often fail to solve the complex real-world data -The need for methods was highlighted.

Furthermore, Van Buren and Groothuis-Odshoorn (2011) pioneered multivariate imputation using chained equations, showing how this method generates multiple data sets to allow more accurate analysis Their study found that multiple imputation outperformed single implant techniques by reducing bias and increasing reliability of estimates findings.

Finally, it is essential to note that selecting an appropriate method for handling missing data depends significantly on the size and characteristics of the dataset, as it directly impacts the statistical power of the analysis, as emphasized by Cohen (1988).

Effects of missing data proportion on statistical analysis accuracy and control methods:

Literature contribution

Research has shown that the proportion of missing data can significantly affect the accuracy of the statistical

analysis, and the methods used to address it Schafer and Graham (2002) showed that when the proportion of missing data is low, it is generally insufficient 5% of which, its effect is generally minimal, and can be controlled by simple methods, e.g. Average or median imputation has been used without significantly affecting the analysis It is shown that small amounts of missing data can be effectively handled by special methods.

When the proportion of missing data falls between 5% and 20%, more advanced methods are needed to reduce bias. Van Buren and Groothuis-Oudshaurn (2011) emphasize the importance of methods such as multiple imputation in such cases to ensure the accuracy of the results, as simpler methods can use information a inadequacy occurs For a higher proportion of missing data, especially over 20%, Graham (2009) highlights May be In such cases, researchers may need to consider more advanced solutions, such as recall or the use of complex statistical techniques to overcome the large differences in the dataset This highlights the importance of understanding the proportion of missing information when choosing the appropriate management strategy it.

Introduction A Practical Introduction: A Practical Approach

In this section we will outline the practical steps taken in this research, focusing on the data used and the processes followed to simulate missing values Data used in in this study were 1,000 complete brain tumors, including various patient populations, medical imaging data and clinical properties Free from honor and values It is, and provides a strong basis for our research.

Data source: The dataset is publicly available on Kaggle, an online platform that stores data of various types across disciplines. It is distributed under the Apache 2.0 License, which allows for unrestricted use for educational and research purposes.

The main purpose of this useful feature is to investigate the effectiveness of imputation methods, especially mean-mean imputation, where data are not available following the Missing Completely at Random (MCAR) procedure . This approach allows us to develop a controlled environment to assess the impact of missing data on statistical analysis results.

Steps of the Practical Methodology

1. Dataset Overview

This dataset gives a complete examination of mind tumor cases, encompassing various patient demographics, clinical imaging information, and clinical attributes. It is designed to facilitate research and development in clinical picture analysis, mainly in the detection and type of brain tumors. The dataset serves as a valuable resource for statistical evaluation, machine gaining knowledge of, and clinical studies, specially within the context of lacking statistics evaluation. The dataset consists of anonymized actual-world instances, making sure patient privateness and confidentiality, and includes the subsequent key attributes:

• Patient ID: A particular identifier for every patient.

• Age: The affected person's age at the time of prognosis.

• Gender: The gender of the affected person (Male/Female).

• Tumor Type: A specific variable classifying tumor kinds (e.G., Meningioma, Glioma, Pituitary Tumor).

• Tumor Location: The anatomical place of the tumor within the brain.

• MRI Images: Scans from multiple MRI modalities including T1-weighted, T2-weighted, and FLAIR.

• Clinical Notes: Detailed medical observations and symptoms stated by way of healthcare providers.

• Treatment Plan: Information on remedy strategies, which include surgical procedure, radiotherapy, chemotherapy, or mixtures thereof.

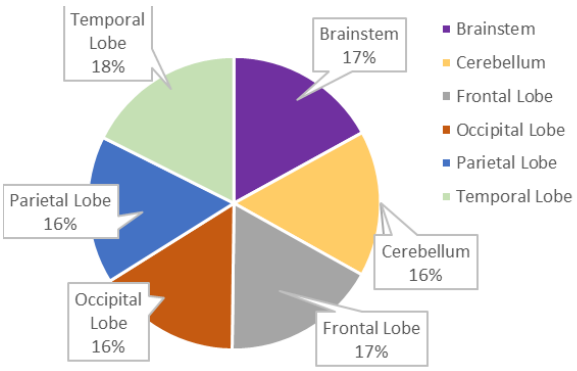
After engaging in an analysis of the records, which served as a foundational step for expertise the characteristics of the facts and could help in next methods related to simulating lacking values and comparing imputation methods, we acquired the subsequent effects:

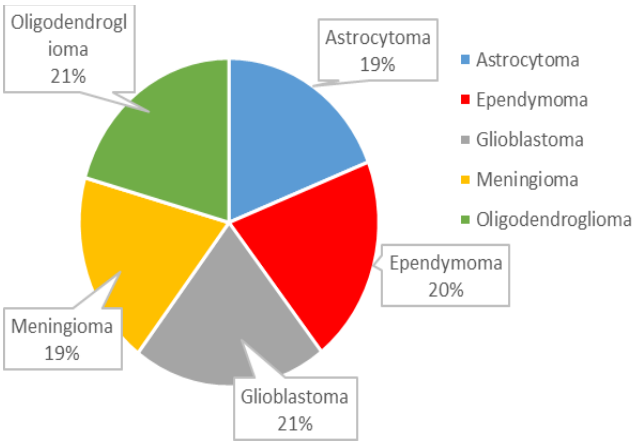
(Descriptive Statistics)

	AVERAGE	MEDIAN	STDV.P	MAX	MIN	VAR.S	SUM
Patient Age	43.519	43	24.99331	89	1	625.2909	43519
Size (cm)	5.2215	5.265	2.825904	10	0.51	7.993727	5221.5

Tumor Type	Ferquency
Astrocytoma	190

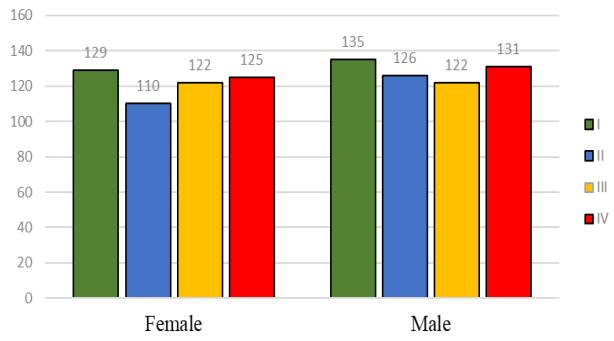
Ependymoma	204
Glioblastoma	210
Meningioma	190
Oligodendroglioma	206
Grand Total	**1000**





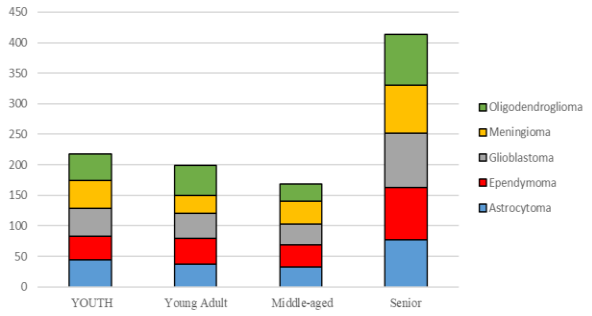
Count of Grade	Column Labels				Grand Total
Row Labels	**I**	**II**	**III**	**IV**	
Female	129	110	122	125	486
Male	135	126	122	131	514
Grand Total	**264**	**236**	**244**	**256**	**1000**

Location	Count of Location
Brainstem	170
Cerebellum	161
Frontal Lobe	171
Occipital Lobe	159
Parietal Lobe	163
Temporal Lobe	176
Grand Total	**1000**



Count of Patient Age	Tumor Type					
Row Labels	**Astrocytoma**	**Ependymoma**	**Glioblastoma**	**Meningioma**	**Oligodendroglioma**	**Grand Total**
YOUTH	44	39	46	45	44	**218**

Young Adult	37	42	41	30	49	**199**
Middle-aged	32	37	34	37	29	**169**
Senior	77	86	89	78	84	**414**
Grand Total	**190**	**204**	**210**	**190**	**206**	**1000**



Regression Statistics	
Multiple R	0.004661
R Square	2.17E-05
Adjusted R Square	-0.00098
Standard Error	2.828703
Observations	1000

2. Simulating Missing Data:

In this phase, we will utilize Excel to systematically generate random missing values across various data fields, adhering to the Missing Completely at Random (MCAR) mechanism. This process will involve specifying the proportion of missing data to be created, with a defined missing data rate of 10%, thus allowing us to explore a variety of scenarios for further analysis. We executed this step specifically for the data related to patient age and tumor size measurements in centimeters (cm). To generate random missing values, we employed the `RAND()` or `RANDBETWEEN()` functions, ensuring a robust simulation of missing data scenarios for our analysis.

3. Application of Imputation Methods:

In this step, we will implement imputation techniques using both mean and median to address the simulated missing values in the dataset. The procedures will detail how to replace the missing data points with the mean or median values of their respective variables

4. Discussion of Results:

In this step, the following table was generated, containing all the results of the data before and after the application of the missing data process using the MCAR method, along with the imputation techniques utilizing mean and median values.

	Patient Age	**MCAR (Patient Age)**	**Size (cm)**	**MCAR (SIZE cm)**
SUM	43519	43039.17	5221.5	5182.941
VAR.S	625.2909	569.8945	7.993727	7.349692
MIN	1	1	0.51	0.51
MAX	89	89	10	10
STDV.P	24.99331	23.86052	2.825904	2.709676
MEDIAN	43	43.03917	5.265	5.182941
AVERAGE	43.519	43.03917	5.2215	5.182941

Examining the results, we find no significant difference between the imputed values for the mean and median, indicating that both methods can successfully recover the missing data. This highlights the importance of imputation emphasizes the role of appropriate methods in statistical analysis, helping to increase the accuracy and quality of the data in future research.

A review of imputation methods

The main objective of this study is to evaluate the effectiveness of imputation methods for dealing with missing values, with a particular focus on mean and median occlusion methods These methods were chosen due to their flexibility and flight effectiveness in dealing with missing information.

Key Findings:

1. Ease of use: Mean and median comparison methods are simple and easy to use, making them suitable for researchers across disciplines, especially when dealing with large data sets.

2. Data recovery: These techniques greatly help in recovering lost data, thereby reducing the negative impact of data loss on statistical results.

3. Improved model accuracy: These imputation techniques can be used to increase the accuracy of statistical analysis models, as more reliable information is obtained from complete data

These findings highlight the usefulness and utility of using mean and median imputation to deal with missing data in statistical analysis.

Conclusions

1. Summary of findings:

• This study examined the effectiveness of the comparison method in a series of brain tumor cases, which included 1,000 complete records. Simulated missing values followed the MCAR (Missing Completely at Random) mechanism, by which The imputation techniques would perform rigorous analysis

• The results showed that the mean and median rating methods restored data integrity and accuracy well, producing similar statistical results. This confirms their reliability in dealing with the challenge of missing data, even for example with a data missing rate of 10%.

2. What the results mean:

• The findings suggest the need for robust imputation techniques to ensure data quality in statistical analysis. The lack of significant differences between the mean imputation results and the derivatives highlights the usefulness of these methods in real-world applications.

3. Strengths of the methods:

• The ease of centralization and centralization, as well as the ease of use, makes these techniques very accessible to researchers across disciplines, especially in large data sets. The ability to successfully recover missing data effectively contributes to the validity of research findings.

4. Limitations:

• While mean and median imputation are useful, they cannot account for underlying data distributions or relationships between variables, which can lead to biased estimates in some cases Future research should explore the limitations of these methods on more complex data sets.

Recommendations

1. Future research directions:

• It is recommended that future research investigate the effectiveness of more advanced seizure techniques, such as multiple seizures or mechanical learning-based methods, and compare their performance with traditional seizure-like methods down to of the decline.

• Researchers should investigate the impact of different data types of missingness and percentages on the use of different imputation methods.

2. Practical applications:

• Professionals in fields such as medical research and clinical research should consider using these attribution techniques in their research to increase the robustness of the findings.

• For better data control, it is advisable to perform a sensitivity analysis to examine how different imputation methods can affect the analysis results.

3. Guidelines for Data Processing:

• Prioritize clear guidelines for managing missing data in research processes. Emphasizing the importance of selecting appropriate comparative methods based on the specific context and characteristics will improve the overall quality of the data.

Following these findings and recommendations, researchers can optimize their data mining techniques, yielding more accurate and reliable results in studies with missing data in the wombs.

REFERENCES

1. Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates
2. Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual Review of Psychology.
3. Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147–177.
4. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3).

5. Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC.
6. Van Buuren, S. (2018). Flexible Imputation of Missing Data. Chapman and Hall/CRC.
7. Enders, C. K. (2010). Applied Missing Data Analysis. Guilford Press. Available on Google Books.
8. Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data. John Wiley & Sons. Available on Google Books.
9. Kaggle. (n.d.). Brain Tumor Cases Dataset. Retrieved from https://www.kaggle.com, under the Apache 2.0 License.