
Data Science & Business Analytics

[Articles](#) [Ebooks](#) [Free Practice Tests](#) [On-demand Webinars](#) [Tutorials](#)

[Home](#) [Resources](#) [Data Science & Business Analytics](#) [Introduction to Data Imputation](#)



Introduction to Data Imputation

By Simplilearn

Last updated on Aug 16, 2023

18579



Table of Contents

[What Is Data Imputation?](#)

[Importance of Data Imputation](#)

[Data Imputation Techniques](#)

[What Is Multiple Imputation?](#)

[Example Of Multiple Imputation](#)

[View More](#)

Imputation in statistics refers to the procedure of using alternative values in place of missing data. It is referred to as "unit imputation" when replacing a data point and as "item imputation" when replacing a constituent of a data point.

Missing information can introduce a significant degree of bias, make processing and analyzing the [data](#) more difficult, and reduce efficiency, which are the three main issues it causes.

Imputation is viewed as an alternative to listwise elimination of cases with missing values since missing data can complicate data analysis.

In other words, most statistical software defaults to dismissing any instance with a missing value when one or more data are absent for a case, which may add bias or impair the generalisability of the results.

By substituting missing information with an estimated value depending on other available information, imputation preserves all cases. The data set can be analyzed using methods used for complete data once all values have been imputed. Scientists have adopted a variety of ideas to explain missing data, but the bulk of them creates bias.

In this article, we will be diving into the world of Data Imputation, discussing its importance and techniques, and also learning about Multiple Imputations.

Become a Data Scientist with Hands-on Training!

Data Scientist Master's Program

[EXPLORE PROGRAM](#)

What Is Data Imputation?

Data imputation is a method for retaining the majority of the dataset's data and information by substituting missing data with a different value. These methods are employed because it would be impractical to remove data from a dataset each time. Additionally, doing so would substantially reduce the dataset's size, raising questions about bias and impairing analysis.

Let us now learn the importance of Data imputation.

Importance of Data Imputation

Now that we learned what Data imputation is, let us see why exactly it is important.



We employ imputation since missing data can lead to the following problems:

- **Distorts Dataset:** Large amounts of missing data can lead to anomalies in the variable distribution, which can change the relative importance of different categories in the dataset.
- **Unable to work with the majority of machine learning-related Python libraries:** When utilizing ML libraries (SkLearn is the most popular), mistakes may occur because there is no automatic handling of these missing data.
- **Impacts on the Final Model:** Missing data may lead to bias in the dataset, which could affect the final model's analysis.
- **Desire to restore the entire dataset:** This typically occurs when we don't want to lose any (or any more) of the data in our dataset because all of it is crucial. Additionally, while the dataset is not very large, eliminating a portion of it could have a substantial effect on the final model.

Since we have explored the importance, we will learn about the various techniques and methods of Data Imputation.

Data Imputation Techniques

After learning about what data imputation is and its importance, we will now learn about some of the various data imputation techniques. These are some of the data imputation techniques that we will be discussing in-depth:

- Next or Previous Value
- K Nearest Neighbors
- Maximum or Minimum Value
- Missing Value Prediction
- Most Frequent Value
- Average or Linear Interpolation
- (Rounded) Mean or Moving Average or Median Value
- Fixed Value

We will be exploring each of these techniques in a detailed manner now.



1. Next or Previous Value

For time-series data or ordered data, there are specific imputation techniques. These techniques take into consideration the dataset's sorted structure, wherein nearby values are likely more comparable than far-off ones. The next or previous value inside the time series is typically substituted for the missing value as part of a common method for imputed incomplete data in the time series. This strategy is effective for both nominal and numerical values.

2. K Nearest Neighbors

The objective is to find the k nearest examples in the data where the value in the relevant feature is not absent and then substitute the value of the feature that occurs most frequently in the group.

3. Maximum or Minimum Value

You can use the minimum or maximum of the range as the replacement cost for missing values if you are aware that the data must fit within a specific range [minimum, maximum] and if you are aware from the process of data collection that the measurement instrument stops recording and the message saturates further than one of such boundaries. For instance, if a price cap has been reached in a financial exchange and the exchange procedure has indeed been halted, the missing price can be substituted with the exchange boundary's minimum value.

4. Missing Value Prediction

Using a machine learning model to determine the final imputation value for characteristic x based on other features is another popular method for single imputation. The model is trained using the values in the remaining columns, and the rows in feature x without missing values are utilized as the training set.

Depending on the type of feature, we can employ any regression or classification model in this situation. In resistance training, the algorithm is used to forecast the most likely value of each missing value in all samples.

A basic imputation approach, such as the mean value, is used to temporarily impute all missing values when there is missing data in more than a feature field. Then, one column's values are restored to missing. After training, the model is used to complete the missing variables. In this manner, an is trained for every feature that has a missing value up until a model can impute of the missing values.

5. Most Frequent Value

The most frequent value in the column is used to replace the missing values in another popular technique that is effective for both nominal and numerical features.

6. Average or Linear Interpolation

The average or linear interpolation, which calculates between the previous and next accessible value and substitutes the missing value, is similar to the previous/next value imputation but only applicable to numerical data. Of course, as with other operations on ordered data, it is crucial to accurately sort the data in advance, for example, in the case of time series data, according to a timestamp.

7. (Rounded) Mean or Moving Average or Median Value

Median, Mean, or rounded mean are further popular imputation techniques for numerical features. The technique, in this instance, replaces the null values with mean, rounded mean, or median values determined for that feature across the whole dataset. It is advised to utilize the median rather than the mean when your dataset has a significant number of outliers.

8. Fixed Value

Fixed value imputation is a universal technique that replaces the null data with a fixed value and is applicable to all data types. You can impute the null values in a survey using "not answered" as an example of using fixed imputation on nominal features.

Since we have explored single imputation, its importance, and its techniques, let us now learn about Multiple imputations.

Become a Data Scientist with Hands-on Training!

Data Scientist Master's Program

EXPLORE PROGRAM



What Is Multiple Imputation?

Single imputation treats an unknown missing value as though it were a true value by substituting a single value for it [Rubin, 1988]. Single imputation overlooks uncertainty as a result, and it almost invariably understates variation. This issue is solved by multiple imputations, which account for both within- and between-imputation uncertainty.

For each missing value, the multiple data imputation approaches generate n suggestions. Each of these values of n is given a plausible value, and n fresh datasets are produced as though a straightforward imputation had taken place in each dataset.

In this fashion, a single table column creates n brand-new sets of data, which are then individually examined using particular techniques. In a subsequent phase, these analyses were combined to produce or consolidate the results of that data set.

The following steps take place in multiple imputations:

Step 1: A collection of n values to also be imputed is created for each attribute in a data set record that is missing a value;

Step 2: Utilizing one of the n replacement ideas produced in the previous item, a statistical analysis is carried out on each data set;

Step 3: A set of results is created by combining the findings of the various analyses.

We will now try to understand this in a better way by looking at an example.

Example Of Multiple Imputation

A perfect example of Multiple Data Imputation is explained below.

Think about a study where some participants' systolic blood pressure information is missing, such as one looking at the relationship between systolic blood pressure and the risk of developing coronary heart disease later on. Age (older patients are more likely to have their systolic blood pressure measured by a doctor), rising body mass index, and a history of smoking all reduce the likelihood that it is missing.




We can use multiple estimations to calculate the overall affiliation between systolic blood pressure and heart disease if we presume that data are missing at random and we have systolic blood pressure information data on a representative sample of people within body mass index, strata of age, coronary heart disease and, smoking.

There is potential for multiple imputations to increase the reliability of medical studies. The user must model the probability of each variable with missing values using the observed data when using the multiple imputation process, though. Multiple imputation results must be modeled carefully and appropriately in order for them to be valid. If at all possible, specialized statistical assistance should be sought before using multiple imputations as a standard procedure that can be used at the touch of a button.

Choose the Right Program

Make a well-informed choice to advance your data science career by exploring our extensive course comparison. We have provided a comprehensive overview of our programs, enabling you to find the ideal one that perfectly matches your goals and aspirations in the dynamic field of data science.

Program Name	Data Scientist Master's Program	Post Graduate Program In D Science
Geo	All Geos	All Geos
University	Simplilearn	Purdue
Course Duration	11 Months	11 Months
Coding Experience Required	Basic	Basic
Skills You Will Learn	10+ skills including data structure, data manipulation, NumPy, Scikit-Learn, Tableau and more	8+ skills including Exploratory Data Analysis Descriptive Statistics, Inferer Statistics, and more
		 Purdue Alumni Associatio

Additional Benefits	Applied Learning via Capstone and 25+ Data Science Projects	Membership Free IIMJobs Pro-Membership 6 months Resume Building Assistance
Cost	\$\$	\$\$\$\$
	Explore Program	Explore Program

Get A Data Analytics Certificate From Simplilearn

In this article, we discussed Data Imputation and its importance. We also discussed some of the main techniques of Data Imputation and also explored multiple imputations along with an example.

To understand this concept in a better way and to implement this while performing [data analysis](#), do consider enrolling in Simplilearn's [Caltech Post Graduate Program in Data Science](#) and take a step towards excelling in your career!

The Exclusive Path to Your Dream Career

Data Science Career Guide

GET YOUR COPY

FAQs

1. What does imputation mean in data?

The replacement of missing or inconsistent data elements with approximated values is known as imputation in data. It is intended for the substituted values to produce a data record that passes

edits.

2. What is data imputation in machine learning?

In Machine Learning, we perform Model-based imputation. Median and mean imputation are two examples of techniques that approximate missing values based on presumptions about the data's distribution that are referred to as "model-based imputation." Alternatively, making assumptions about the link between the target y variable and auxiliary variables (or x variables) to anticipate missing values.

3. What are the data imputation techniques?

Some of the various data imputation techniques are:

- Next or Previous Value
- K Nearest Neighbors
- Maximum or Minimum Value
- Missing Value Prediction
- Most Frequent Value
- Average or Linear Interpolation
- (Rounded) Mean or Moving Average or Median Value
- Fixed Value

4. When should you impute data?

Imputation generates plausible hypotheses for lacking data. It works best when there are a few missing data points.

5. Why is data imputation important?

By substituting missing data with an average worth based on some other available information, imputation preserves all cases. The data set can be analyzed using methods used for complete data once all values have been imputed.

6. How do you impute missing values in data?



The statistics (mean, median, or most common) of each row where the missing values are present can be used to impute missing values, or they can be replaced with a constant value.

7. What is the difference between interpolation and imputation?

While imputation replaces missing data for the column's mean, interpolation is a sort of estimation that creates data points within the range of a discrete set of existing data points.

Find our Caltech Post Graduate Program in Data Science Online Bootcamp in top cities:

Name	Date	Place
Caltech Post Graduate Program in Data Science	Cohort starts on 18th Dec 2023, Weekend batch	Your City
Caltech Post Graduate Program in Data Science	Cohort starts on 9th Jan 2024, Weekend batch	Your City

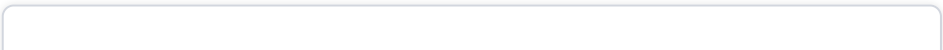
About the Author



Simplilearn

Simplilearn is one of the world’s leading providers of online training for Digital Marketing, Cloud Computing, Project Management, Data Science, IT, Software Development, and many other emerging ...

Recommended Programs



Caltech Post Graduate Program in Data Science

637 Learners

Lifetime
Access*

Data Scientist

50291 Learners

Lifetime
Access*

*Lifetime access to high-quality, self-paced e-learning content.

[Explore Category](#)

NEXT ARTICLE

What a DevOps Post Graduate Certification From Caltech CTME and Simplilearn Means for You

By Simplilearn

1561

Feb 7, 2023

Recommended Resources

Data Science Career Guide: A
Comprehensiv...

Data Science Grad
Programs to Launc



Disclaimer

PMP, PMI, PMBOK, CAPM, PgMP, PfMP, ACP, PBA, RMP, SP, and OPM3 are registered marks of the Project Management Institute, Inc.

