Missing Values and Absent Values

Alassane SOMA

October 23, 2024

Contents

1 Is there a difference between missing values and absent values?

The answer is ”Yes.” In statistics and data analysis, a missing value and an absent value may seem similar, but they have important distinctions depending on the context in which they are used:

1.1 Missing Value

A missing value refers to expected data in a dataset that is unavailable or was not collected for some reason. This means that an observation was supposed to be made or recorded at a specific point, but the information is missing. It is typically represented by special symbols (such as ‘NA‘, ‘NaN‘, or ‘.‘ depending on the software used) [1].

Example: In a survey, if a question is posed to all participants but one person does not respond, the answer to that question is considered a missing value.

1.2 Absent Value

An absent value refers to data that is not expected or does not exist in a given context. This means the absence of data is intentional or natural because it is not relevant for that specific observation [2].

Example: In a survey that only applies to a specific subgroup, such as women, men would not have answers to certain questions (e.g., questions related to pregnancy). In this case, the data is absent for men, as the information is not applicable to them.

1.3 Key Differences

- **Missing Value:** Data that should have been present but is not (an unexpected absence).
- **Absent Value:** Data that is not applicable or expected for a certain subset of observations (a planned absence).

In summary, a missing value reflects a flaw in data collection, while an absent value may result from a situation where the data is simply irrelevant [3].

2 How to handle missing values?

When dealing with data, missing values can cause problems if not handled correctly, as they can distort results or make certain analyses impossible. Here are some common approaches[1] to address missing values before performing an analysis:

[1] Several methods exist, but here are a few popular ones.

2.1 Remove Observations with Missing Values (Listwise Deletion)

This method involves completely excluding observations that contain missing values. It is simple to implement but may result in significant data loss if many observations have missing values.

- **Advantages:** Simple to apply, maintains statistical validity.
- **Disadvantages:** May reduce sample size, leading to loss of valuable information and decreased statistical power [4].

2.2 Imputation of Missing Values

Imputation involves replacing missing values with plausible values based on the available information. There are several methods of imputation, such as:

2.2.1 Mean/Median/Mode Imputation

Replace missing values with the mean (for continuous variables), median, or mode (for categorical variables) of the other observations [5].

- **Advantages:** Simple and quick.
- **Disadvantages:** Can bias results by reducing variability in the data and distorting relationships between variables.

2.2.2 Regression Imputation

Use regression models to predict missing values based on other variables in the dataset [6].

- **Advantages:** More accurate approach, takes into account relationships between variables.
- **Disadvantages:** More complex, may underestimate variance.

2.2.3 K-Nearest Neighbors (K-NN) Imputation:

This method imputes missing values using the mean or median of the K closest observations (in terms of distance) in the dataset [7].

- **Advantages:** Uses local information, flexible.
- **Disadvantages:** Requires a good estimation of nearest neighbors, computationally expensive.

2.2.4 Multiple Imputation

Multiple imputation involves generating several imputed datasets, analyzing them separately, and then combining the results to obtain a final estimate.

- **Advantages:** Statistically rigorous approach that accounts for uncertainty in missing values.
- **Disadvantages:** Complex to implement, requires specialized software.

2.3 Use Robust Statistical Methods for Missing Values

Some analysis methods are designed to work with data that contains missing values without requiring imputation. For example, maximum likelihood regression models or structural equation models (SEM) can directly handle missing data.

- **Advantages:** No need for additional data manipulation.
- **Disadvantages:** Not suitable for all types of analysis.

2.4 Predict Missing Values Using Advanced Algorithms

In more advanced contexts, machine learning algorithms like decision trees, neural networks, or random forests can be used to predict missing values based on complex relationships between variables.

- **Advantages:** Can capture complex relationships between variables.
- **Disadvantages:** Complex to implement, requires significant computational resources.

2.5 Create a ”Missing” Category (for Categorical Variables)

For categorical variables, another solution is to create a new ”Missing” category to code observations that have missing values.

- **Advantages:** Simple and transparent.
- **Disadvantages:** Can be biased if the missing values are not random.

2.6 Models for Non-Random Missing Data

If the missing data is not random (i.e., there is a specific reason why it is missing), it may be useful to use models designed for missing data. Bias estimation models are examples of approaches suited for these situations.

2.7 Choosing the Method

The choice of method depends on several factors:

- **Percentage of Missing Data:** If the missing data represents a small proportion of the observations, removing rows may be appropriate. If the percentage is high, imputation or more sophisticated methods may be necessary.
- **Type of Missing Data:** Missing data can be completely random (MCAR), conditionally random (MAR), or non-random (MNAR), and this influences the choice of treatment method [8].
- **Objective of the Analysis:** Some analyses are more sensitive to missing values than others. For example, imputation methods may introduce bias in certain predictive analyses.

Retain: Missing values must be handled carefully to avoid bias and ensure the quality of results. The method chosen depends on the context, the nature of the data, and the objective of the analysis. It is essential to understand why the data is missing before selecting a treatment method.

References

[1] Scribbr: Missing Data

[2] Statistics By Jim: Missing Data Basics

[3] Baeldung: Missing vs Sparse Data

[4] YData: Understanding Missing Data Mechanisms

[5] IBM: Handling Missing Values

[6] DataCamp: Techniques to Handle Missing Data Values

[7] Stats with R: Handling Missing Data

[8] Stef van Buuren: Missing Completely at Random (MCAR)