Univariate Imputation- Mean/Median (Numerical Data)

06 July 2023 21:11

2. Imputing (Replacing) Missing Values:

2nd method of Handling Massing Values as **First one** is **Removing/Deleting the missing values** we have already covered.

Univariate Imputation:

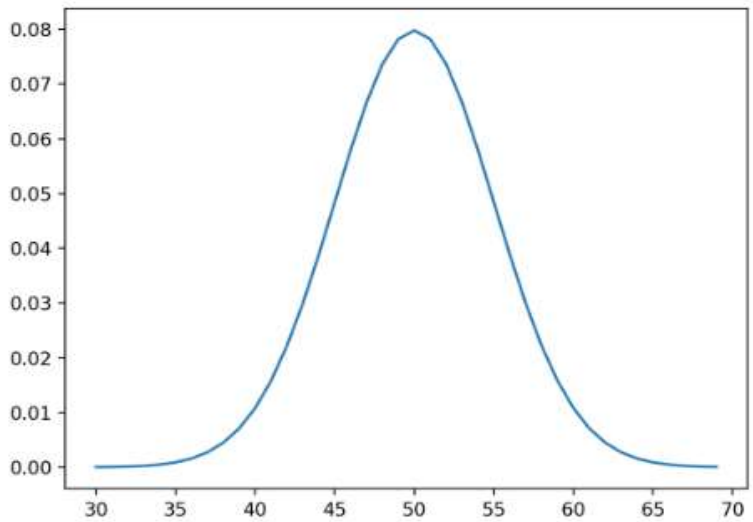
This is a method of **replacing missing values** in a feature with a value that is **estimated from the non-missing values in the same feature**.
This estimated value depends on the type of feature containing Missing Values.

Numerical Feature:- Mean, Median, End of Distribution Value, Any Random Value etc.

Univariate Imputation on Numerical Features:

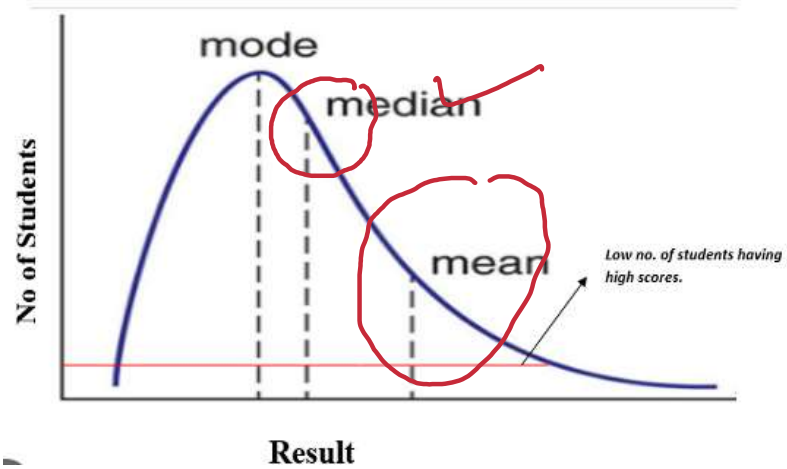
1. **Imputing Mean/Median:**
   It involves replacing the missing values in a feature with Mean/Median of that feature.

a. **Mean**:- when data is **normally distributed**, then we can use Mean for imputation.



The **mean** is a **measure of central tendency** that represents the average value of the feature.
Imputing missing values with the mean preserves the overall average and does not significantly distort the distribution.

b. **Median**:- when data is **skewed**, then we can use Median for imputation.



Median imputation for skewed data:
When the distribution of the feature is skewed or contains outliers, the mean may be influenced by extreme values, making it less representative of the typical values.
In such cases, the **median**, which **represents the middle value of the sorted feature**, can be a more robust measure of central tendency.
Imputing missing values with the median is less sensitive to extreme values and can provide a more accurate estimate of the central value for skewed distributions.

Advantages of Mean/Median Imputation:

1. **Simple and Easy to implement**.
2. **Gives good result** when missing values are less than 5%.

Disadvantages of Mean/Median Imputation:.

1. After imputation by central value, the System starts detecting some feature values as **outliers** because of increase in number of values closer to mean which consequently make far off values get detected as outliers.
2. **Changes the Covariance and Correlation** of the feature with other features.

When to use:

1. When missing data is **MCAR** (Missing Completely At Random).
2. When missing values are **less than 5%** of the feature values.