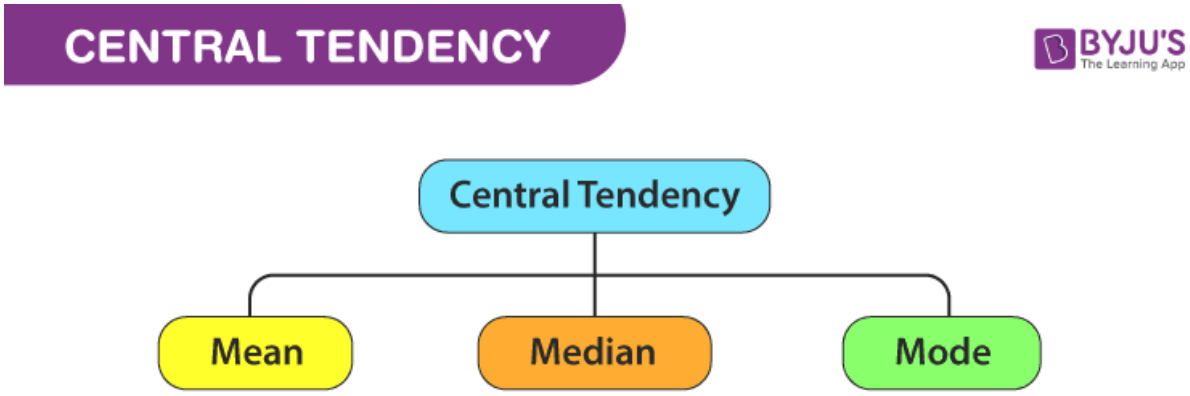
Central Tendency

Definition

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.

The central tendency of the dataset can be found out using the three important measures namely mean, median and mode.



1. Mean

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean. Some other measures of mean used to find the central tendency are as follows:

1. Arithmetic Mean

- **Definition**: The sum of all values divided by the number of values.
- **When to Use**:
    - When dealing with interval or ratio data.
    - When data is symmetrically distributed without extreme outliers.
    - Commonly used for general average calculations in various fields.

2. Geometric Mean

- **Definition**: The nth root of the product of all values (where n is the number of values).

$$\text{Geometric mean (GM)} = \sqrt[n]{(x_1)(x_2)\ldots(x_n)}$$
$$\text{Log (GM)} = \frac{\Sigma(\log x)}{n}$$

- **When to Use**:
    - When dealing with multiplicative processes or growth rates (e.g., population growth, investment returns).

- When data spans several orders of magnitude (e.g., economic data with large ranges).
- Suitable for datasets with skewed distributions or when comparing different rates.

3. Harmonic Mean

- **Definition**: The reciprocal of the arithmetic mean of the reciprocals of the data values.

$$\text{Harmonic mean (HM)} = \frac{1}{\frac{\Sigma(1/x)}{n}} = \frac{n}{\Sigma(1/x)}$$

- **When to Use**:
  - When dealing with rates or ratios (e.g., average speed, price per unit).
  - When the dataset involves variables that are rates of change.
  - Best used when the smaller values are more significant or have a greater imact on the mean.

4. Weighted Mean

- **Definition**: The sum of all values multiplied by their respective weights, divided by the sum of the weights.

$$\text{Weighted mean} = \frac{\Sigma w\bar{x}}{\Sigma w}$$

- **When to Use**:
  - When different values have different levels of importance or frequency.
  - Common in scenarios like calculating GPA, where different courses have different credit values.
  - Useful when combining data from different sources or with varying significance.

Advantages:

Simplicity: Easy to compute and understand.

Overall Summary: Provides a central value that represents the dataset.

Basis for Other Analyses: Useful for further statistical analysis and comparisons.

Balance Point: Minimizes the sum of squared deviations from the mean, making it a good measure of central tendency for normally distributed data.

Standardization: Helps in standardizing and comparing different datasets.

Disadvantages:

Sensitivity to Outliers: Can be skewed by extremely high or low values.

Not Always Representative: May not reflect the typical value in skewed distributions.

Not Suitable for All Data Types: Not appropriate for categorical or ordinal data.

Loss of Information: Does not provide information about the distribution of data or variability.

2. Median

Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.

Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2

Advantages:

Robust to Outliers: Not affected by extremely high or low values, making it a better measure of central tendency for skewed distributions.

Represents the Middle Value: Provides the value that separates the higher half from the lower half of the dataset, which can be more representative in skewed distributions.

Simple to Calculate: Easy to determine, especially for small datasets.

Useful for Ordinal Data: Can be used with ordinal data, where mean is not appropriate.

Reflects Distribution: Provides a good summary of data distribution, especially when data is not symmetrically distributed.

Disadvantages:

Not Sensitive to All Data: Does not take into account the magnitude of values, only their order.

Less Informative: Doesn't provide information about the spread or variability of the data.

Complex for Large Datasets: Determining the median can be cumbersome with large datasets, especially if they are not sorted.

Not Affected by All Distributions: May not reflect the central tendency if the dataset has multiple peaks or is multimodal.

3. Mode

The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5

Advantages:

Simplicity: Easy to identify and understand, even with categorical data.

Useful for Categorical Data: Can be used to determine the most common category or value in a dataset.

Highlights Popular Trends: Shows the most frequently occurring value, which can be useful for identifying trends or popular items.

No Assumption About Distribution: Doesn't require assumptions about the data's distribution, making it versatile for different types of data.

Applicable to Discrete Data: Can be used with discrete data where other measures like mean might not be suitable.

Disadvantages:

Not Unique: A dataset can have more than one mode (bimodal or multimodal), or no mode at all if all values occur with the same frequency.

Not Always Representative: The mode may not provide a good summary of the data if it's not the most central value.

Not Suitable for All Data Types: Less useful for continuous data where values can take on an infinite range.

Limited Insight: Provides only information about the most frequent value without detailing the overall distribution or variability.

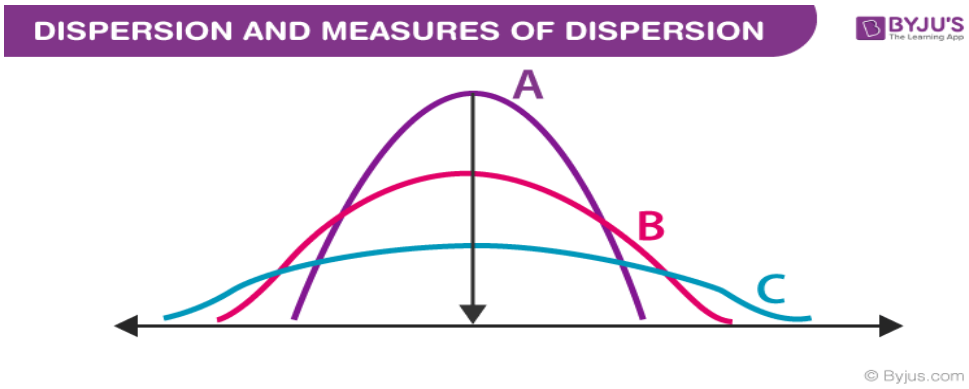
Based on the properties of the data, the measures of central tendency are selected.

- If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset.
- If you have skewed distribution, the best measure of finding the central tendency is the median.
- If you have the original data, then both the median and mode are the best choice of measuring the central tendency.
- If you have categorical data, the mode is the best choice to find the central tendency.

- The **mean** is the most sensitive to changes in scale and location.
- The **median** is sensitive to location changes but not scale changes.
- The **mode** is least sensitive to changes in scale and location.

Dispersion and Measures of Dispersion

Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.



Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

Absolute Measure of Dispersion

An absolute measure of dispersion contains the same unit as the original data set. The absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations

The types of absolute measures of dispersion are:

1. Range

When to Use:

- **Initial Overview**: When you need a quick, simple measure of the overall spread of the data.
- **Small Datasets**: When dealing with smaller datasets where extreme values or outliers have a more pronounced effect.
- **Comparing Datasets**: To get a basic sense of how spread out two datasets are, especially if they have similar ranges.

Advantages:

- Simple to compute.
- Provides a basic measure of variability.

Disadvantages:

- Highly sensitive to outliers.
- Does not account for the distribution of data between the extremes.

2. Interquartile Range (IQR)

When to Use:

- **Robust Measure**: When you need a measure of dispersion that is less affected by outliers or extreme values.
- **Box Plots**: When creating box plots to visualize the spread of the middle 50% of data.
- **Non-Normal Distributions**: When dealing with skewed distributions or non-normal datasets.

Advantages:

- Less affected by outliers compared to the range.
- Focuses on the middle 50% of the data, giving a more robust measure of central spread.

Disadvantages:

- Does not use all data points, ignoring the lowest 25% and highest 25%.

3. Variance

When to Use:

- **Detailed Analysis**: When you need a detailed measure of how data points deviate from the mean.
- **Statistical Modeling**: In statistical analyses that require variance for calculations, such as hypothesis testing or regression analysis.
- **Comparing Variability**: To compare the spread of different datasets or distributions.

Advantages:

- Comprehensive measure considering all data points.
- Essential for many statistical tests and models.

Disadvantages:

- Sensitive to outliers due to squaring deviations.
- Results are in squared units, which can be less intuitive.

4. Standard Deviation

When to Use:

- **Interpretability**: When you need a measure of spread in the same units as the data for easier interpretation.
- **Normal Distributions**: When the data is approximately normally distributed, where standard deviation has clear interpretive value (68-95-99.7 rule).
- **Advanced Analysis**: When performing statistical analyses or creating confidence intervals.

Advantages:

- Easy to interpret in the context of the original data units.
- Useful for many statistical applications and tests.

Disadvantages:

- Sensitive to outliers, similar to variance.
- May be less robust in non-normal distributions.

5. Mean Absolute Deviation (MAD)

When to Use:

- **Robust Spread Measure**: When you want a measure of dispersion that is less sensitive to outliers than the standard deviation.
- **Non-Normal Data**: For data that is not normally distributed or has significant outliers.

Advantages:

- Less sensitive to outliers compared to standard deviation.
- In the same units as the data, making it easier to interpret.

Disadvantages:

- Less commonly used in statistical analyses compared to variance and standard deviation.
- Does not have as many desirable mathematical properties for advanced statistical methods.

Summary

- **Range**: Quick and simple; useful for initial assessments.
- **Interquartile Range (IQR)**: Robust to outliers; good for skewed data.
  Upper limit = Q3 + (1.5)IQR
  Lower limit = Q1 - (1.5)IQR
- **Variance**: Detailed measure; essential for many statistical analyses.
- **Standard Deviation**: Intuitive and useful for data in normal distributions; widely used.
- **Mean Absolute Deviation (MAD)**: Robust to outliers; simpler than standard deviation but less commonly used.

Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

1. Co-efficient of Range
2. Co-efficient of Variation
3. Co-efficient of Standard Deviation
4. Co-efficient of Quartile Deviation
5. Co-efficient of Mean Deviation