Mean, median, and mode are measures of central tendency in statistics, which describe the typical or central value of a dataset.

Mean is calculated by adding all the values in a dataset and then dividing by the number of values. For example, consider the following dataset of exam scores: 70, 85, 90, 75, 80. To find the mean, we add up all the scores and divide by the total number of scores: (70 + 85 + 90 + 75 + 80) / 5 = 80. This means the mean score in this dataset is 80.

Median is the middle value of a dataset when the values are arranged in ascending or Descending order. If the number of values in the dataset is even, the median is calculated by taking the average of the two middle values. For example, in the same dataset of exam scores (70, 85, 90, 75, 80), we arrange the values in order from smallest to largest: 70, 75, 80, 85, 90. The median score is the middle value, which is 80 in this case.

Mode is the value that appears most frequently in a dataset. If there are multiple values that appear with the same frequency, the dataset is considered to have multiple modes. For example, consider the following dataset of shoe sizes for a group of people: 8, 9, 9, 10, 10, 10, 11. The mode in this dataset is 10, since it appears most frequently.

All three measures provide useful information about a dataset, but they each have their own strengths and weaknesses. Mean is useful when the dataset is normally distributed (i.e. the values are evenly distributed around the center) and when there are no extreme values that would skew the results. Median is more robust than mean when the dataset contains outliers or extreme values. Mode is useful when you want to know the most frequently occurring value in a dataset, which can be important for certain types of data analysis.

Sure, let me give you an example to illustrate the point:

Suppose we have a set of salaries for a small company with five employees:

{30,000, 40,000, 45,000, 55,000, 500,000}

The mean salary for this group is calculated by adding up all the salaries and dividing by the number of employees, which gives us:

(30,000 + 40,000 + 45,000 + 55,000 + 500,000) / 5 = 134,000

The mean salary for this group is $134,000, which is heavily influenced by the single high salary of $500,000.

On the other hand, the median salary is simply the middle value when the salaries are arranged in order:

30,000, 40,000, 45,000, 55,000, 500,000

The median salary in this case is $45,000, which is a more representative measure of central tendency because it is not affected by the extreme value of $500,000.

Overall, the median is a valuable tool for summarizing data in a way that is less sensitive to outliers than the mean, making it a more robust measure of central tendency.

Mode : it is useful because it gives the most commonly occurring value in a set of data. It can be useful to identify the most popular or frequent category or value in a dataset. For example, the mode of the favorite color of a group of people can tell you which color is the most popular among that group.

The ideal range of skewness depends on the nature of the data and the purpose of the analysis. However, in general, a skewness value between -1 and 1 is considered as an acceptable range, where a value of 0 indicates a perfectly symmetrical distribution.

A positively skewed distribution (skewness > 0) indicates that the tail of the distribution is longer on the positive side and the mean is greater than the median. In this case, the median is a better measure of central tendency.

On the other hand, a negatively skewed distribution (skewness < 0) indicates that the tail of the distribution is longer on the negative side and the mean is less than the median. In this case, the mean is a better measure of central tendency.

However, it's important to note that skewness is just one aspect of the distribution, and it's important to consider other factors such as the sample size, the presence of outliers, and the purpose of the analysis when selecting the appropriate measure of central tendency.

What is the Data ?

Ans : collection of information based on requirement

Once you’ve recorded some data, you’ll be able to analyze the data for any information they can provide.

business analytics : The use of data analysis to aid in business decision making.

Predictive analytics : The use of data analysis designed to form predictions about future, or unknown, events or outcomes.

business strategy : A plan of action designed by a business practitioner to achieve a business objective.

Business objective : profits, customer satisfaction

Creation of new customers, improving productivity

Best utilization of resources.

Without data taking decisions are assumptions only

With data : it allows us to make evidence based assessment

three key features that we should identify before attempting to analyze any dataset: whether the data are structured or not, the unit of observation, and the datagenerating process.

structured data Data with well-defined units of observation for which corresponding information is identifiable; they are the data that come in a spreadsheet format.

Or

Structured data is quantitative data in the form of numbers and values.

Quantity must be same. Eg 5025 , 5,025 both are not same so it is not structured data.

YEAR SALES

2012 5062

2014 2365

2016 2564 structured data

Unstructered Data : Any data that cannot be classified as structured.

It is qualitative data in the form of text files, audio files , images.

unit of observation : The entity for which information has been collected.

Or

What is your unit of observation?

A unit of observation is **the item (or items) that you actually observe, measure, or collect in the course of trying to learn something about your unit of analysis**.

Independent variable is unit of observation.

The four main groupings of units of observation are cross-sectional data, pooled cross-sectional data, time-series data, and panel data.

cross-sectional data : Data that provide a snapshot of information at one fixed point in time

who , what ,when , where 4Ws.

pooled cross-sectional data : The result of two or more unrelated cross-sectional datasets being combined into one dataset.

time-series data Data that exhibit only variation in time

panel data : The same cross-sectional units over multiple points in time.