# What is Statistics

Statistics is a branch of mathematics that involves collecting, analysing, interpreting, and presenting data. It provides tools and methods to understand and make sense of large amounts of data and to draw conclusions and make decisions based on the data.

In practice, statistics is used in a wide range of fields, such as business, economics, social sciences, medicine, and engineering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions, and make forecasts and predictions.

Examples:

1. Business - Data Analysis(Identifying customer behavior) and Demand Forecasting
2. Medical - Identify efficacy of new medicines(Clinical trials), Identifying risk factor for diseases(Epidemiology)
3. Government & Politics - Conducting surveys, Polling
4. Environmental Science - Climate research

# Types of Statistics

## 1. Descriptive:

Descriptive statistics deals with the collection, organization, analysis, interpretation, and presentation of data. It focuses on summarizing and describing the main features of a set of data, without making inferences or predictions about the larger population.

## 2. Inferential:

Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables.

# Population Vs Sample:

Population refers to the entire group of individuals or objects that we are interested in studying. It is the complete set of observations that we want to make inferences about. For example, the population might be all the students in a particular school or all the cars in a particular city.

A sample, on the other hand, is a subset of the population. It is a smaller group of individuals or objects that we select from the population to study. Samples are used to estimate characteristics of the population, such as the mean or the proportion with a certain attribute. For example, we might randomly select 100 students.

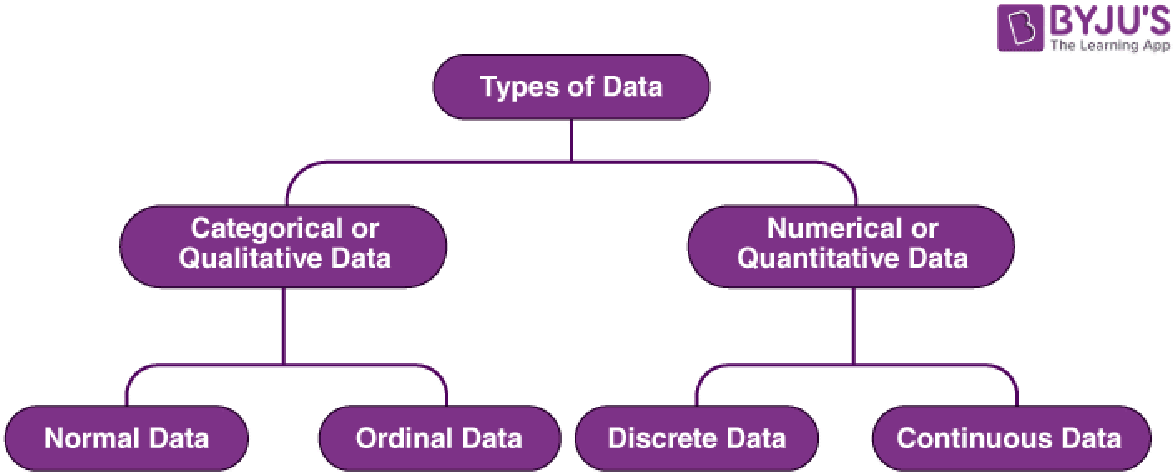
### Examples:

1. All cricket fans vs fans who were present in the stadium
2. All students vs who visit college for lectures

Things to be careful about which creating samples

1. Sample Size
2. Random
3. Representative

## Types of Data

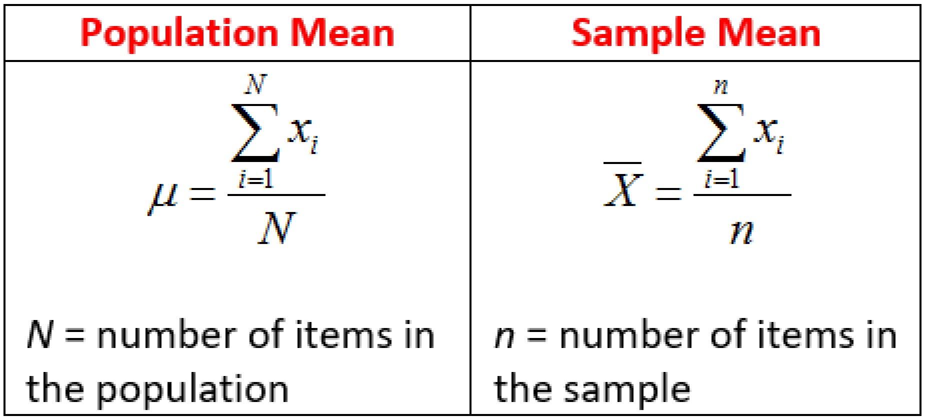


## Measure of Central Tendency:

A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.

### 1. Mean:

The mean is the sum of all values in the dataset divided by the number of values.



| Population Mean | Sample Mean |
|---|---|
| $$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$ $N$ = number of items in the population | $$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$ $n$ = number of items in the sample |

## 2. Median:

The median is the middle value in the dataset when the data is arranged in order.

## 3. Mode:

The mode is the value that appears most frequently in the dataset.

**Note:** Mode is usefull with category data or with discrete data to check frequent values.

## Trimmed Mean:

A trimmed mean is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called the trimming percentage.

Example :

Values:
20,22,23,25,28,30,32,35,50,80

Actual mean : 34.5

Trimmed Values:
25,28,30,32,35

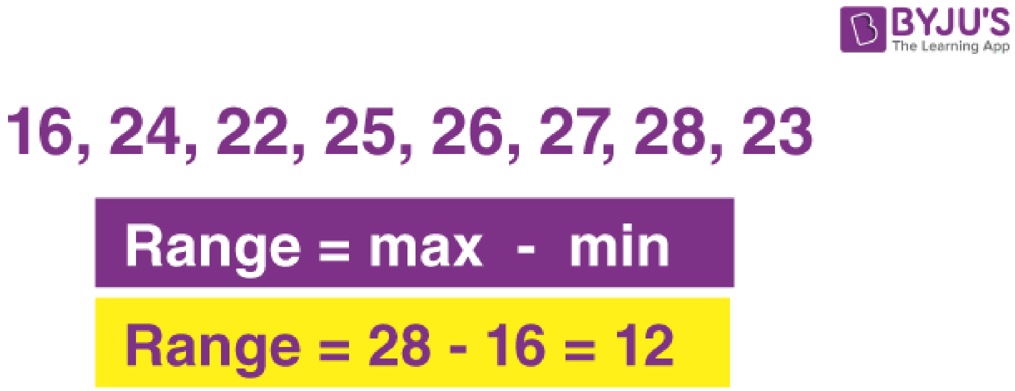
Actual mean : 30

# Measure of Dispersion:

A measure of dispersion is a statistical measure that describes the spread or variability of a dataset. It provides information about how the data is distributed around the central tendency (mean, median or mode) of the dataset.

## 1. Range:

The range is the difference between the maximum and minimum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.



BYJU'S
The Learning App

16, 24, 22, 25, 26, 27, 28, 23

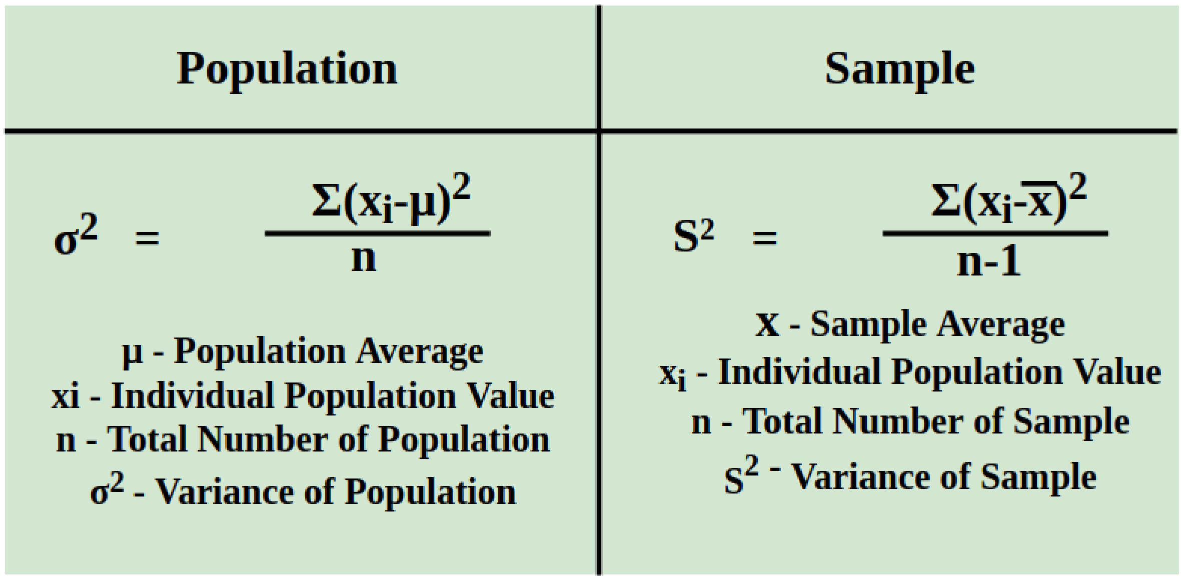
Range = max - min

Range = 28 - 16 = 12

## 2. Variance:

The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

| x | (x-mean) | (x-mean)^2 |
|---|---|---|
| 3 | 3 - 3 = 0 | 0 |
| 2 | 2 - 3 = -1 | 1 |
| 1 | 1 - 3 = -2 | 4 |
| 5 | 5 - 3 = 2 | 4 |
| 4 | 4 - 3 = 1 | 1 |

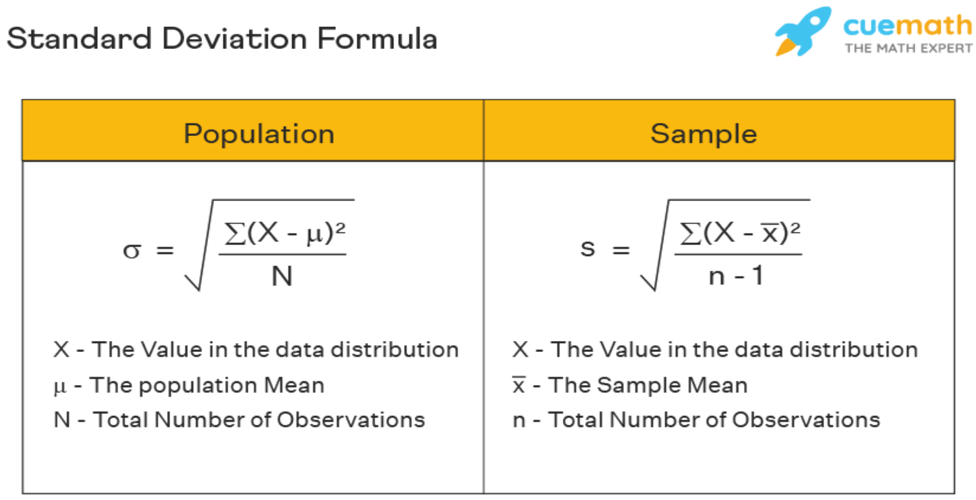


| Population | Sample |
|---|---|
| $\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{n}$ | $S^2 = \frac{\Sigma(x_i-\overline{x})^2}{n-1}$ |
| $\mu$ - Population Average<br>xi - Individual Population Value<br>n - Total Number of Population<br>$\sigma^2$ - Variance of Population | x - Sample Average<br>$x_i$ - Individual Population Value<br>n - Total Number of Sample<br>$S^2$ - Variance of Sample |

## 3. Standard Deviation:

The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

| x | (x-mean)^2 |
|---|---|
| 15 | (15 -14)^2 = 1 |
| 17 | (17 - 14)^2 = 9 |
| 13 | (13 - 14)^2 = 1 |
| 11 | (11 - 14)^2 = 9 |



Standard Deviation Formula

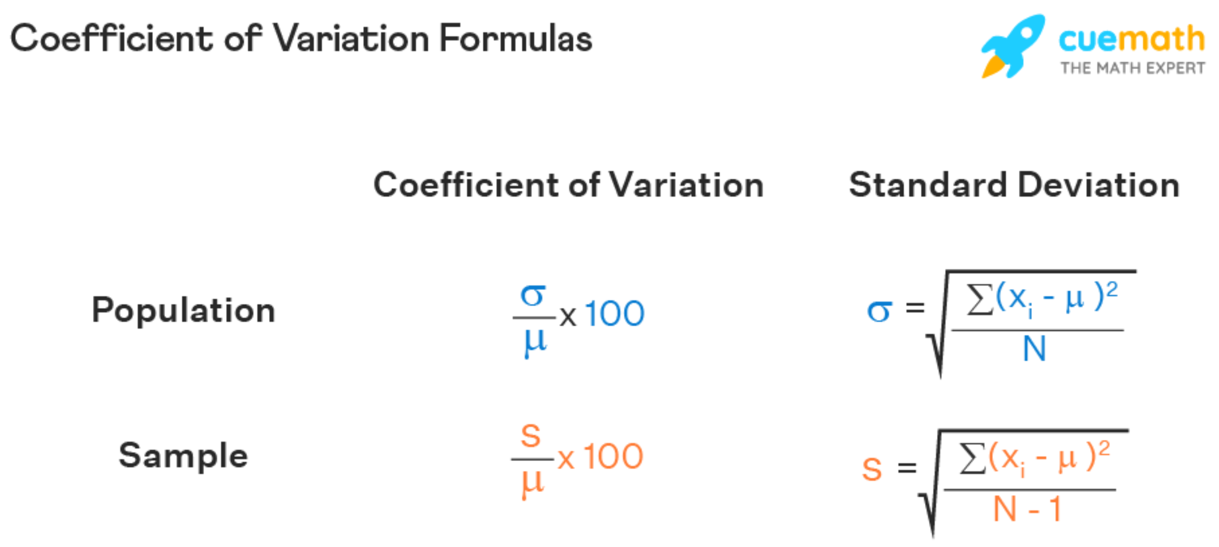
| Population | Sample |
|---|---|
| $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$ | $s = \sqrt{\frac{\Sigma(X-\overline{x})^2}{n-1}}$ |
| X - The Value in the data distribution<br>$\mu$ - The population Mean<br>N - Total Number of Observations | X - The Value in the data distribution<br>$\overline{x}$ - The Sample Mean<br>n - Total Number of Observations |

## Coefficient of Variation:

Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in fields such as biology, chemistry, and engineering.

The coefficient of variation (CV) is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is a dimensionless quantity that is expressed as a percentage.

The formula for calculating the coefficient of variation is:



Coefficient of Variation Formulas

cuemath
THE MATH EXPERT

| | Coefficient of Variation | Standard Deviation |
|---|---|---|
| Population | $\frac{\sigma}{\mu} \text{x}\, 100$ | $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ |
| Sample | $\frac{s}{\mu} \text{x}\, 100$ | $s = \sqrt{\frac{\sum(x_i - \mu)^2}{N - 1}}$ |

## Quantiles and Percentiles :

Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.

Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles used in statistical analysis, including:

1. Quartiles: Divide the data into four equal parts, Q1 (25th percentile) Q2 (50th percentile or median), and Q3 (75th percentile).
2. Deciles: Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile), ..., D9 (90th percentile).
3. Percentiles: Divide the data into 100 equal parts, P1 (1st percentile), P2 (2nd percentile), ..., P99 (99th percentile).
4. Quintiles: Divides the data into 5 equal parts

## Things to remember while calculating these measures:

1. Data should be sorted from low to high
2. You are basically finding the location of an observation
3. They are not actual values in the data
4. All other tiles can be easily derived from Percentiles

## Percentile :

A percentile is a statistical measure that represents the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

Formula to calculate the percentile value:

$$\text{PL} = \frac{P}{100}(N + 1)$$

PL = the desired percentile value location
N = the total number of observations in the dataset
p = the percentile rank (expressed as a percentage)

Example: Find the 75th percentile score from the below data :

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

# 5 number summary:

The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

1. Minimum value:
The smallest value in the dataset.

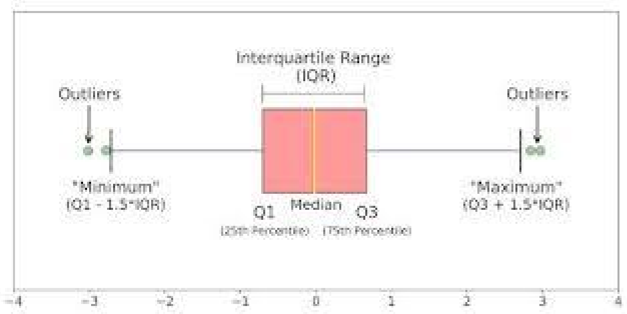
2. First quartile (Q1): The value that separates the lowest 25% of the data from the rest of the dataset.

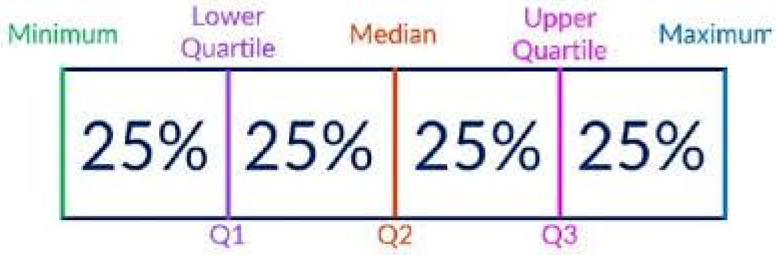
3. Median (Q2): The value that separates the lowest 50% from the highest 50% of the data.

4. Third quartile (Q3): The value that separates the lowest 75% of the data from the highest 25% of the data.

5. Maximum value: The largest value in the dataset.

The five-number summary is often represented visually using a box plot, which displays the range of the dataset, the median, and the quartiles. The five-number summary is a useful way to quickly summarize the central tendency, variability, and distribution of a dataset.





## Interquartile Range

The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.