**-: Statistics Basics:-**

**Agenda:-**

* Skills and tools required for Data Analysis
* Challenges of the Data Analytics Process
* Data visualization techniques
* Hypothesis testing to analyze data
* Exploratory Data Analysis technique
* Data analytics process and its steps

**Statistical and non-statistical analysis**:-

|  |  |
| --- | --- |
| **Statistical** | **Non-statistical** |
| Statistical analysis is scientific | It is not scientific |
| It is based on number or statistical values | It is based on very generic information |
| It is useful in providing complete insight to the data. | It is exclusive of statistical and quantitative analysis |

Although, both forms provide results, quantitative analysis gives better insight to data and clearer picture. That is why statistical analysis is important for business.

**Major categories of Statistics**:-

* Inferential method
* Descriptive method

For Example, for a group of people, we want to analyse the height. There could be two ways for it:-

1) Find the range for tall, medium and small. This accounts for *inferential statistical method*.

2) Find the tallest, smallest and the average among the group. This accounts for *descriptive statistical method*.

Population

Sample

**Measure of central tendency and measure of dispersion**:-

In [statistics](https://en.wikipedia.org/wiki/Statistics), a central tendency (or measure of central tendency) is a central or typical value for a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution). It may also be called a centre or location of the distribution. Colloquially, measures of central tendency are often called [averages](https://en.wikipedia.org/wiki/Averages). The most common measures of central tendency are the [arithmetic ***mean***](https://en.wikipedia.org/wiki/Arithmetic_mean), the [***median***](https://en.wikipedia.org/wiki/Median) and the [***mode***](https://en.wikipedia.org/wiki/Mode_(statistics)).

[**Arithmetic mean**](https://en.wikipedia.org/wiki/Arithmetic_mean)**or simply, mean**

The sum of all measurements divided by the number of observations in the data set.

[**Median**](https://en.wikipedia.org/wiki/Median)

The middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for [ordinal data](https://en.wikipedia.org/wiki/Level_of_measurement#Ordinal_scale), in which values are ranked relative to each other but are not measured absolutely.

[**Mode**](https://en.wikipedia.org/wiki/Mode_(statistics))

The most frequent value in the data set. This is the only central tendency measure that can be used with [nominal data](https://en.wikipedia.org/wiki/Level_of_measurement#Nominal_scale), which have purely qualitative category assignments.

In [statistics](https://en.wikipedia.org/wiki/Statistics), ***dispersion*** (also called ***variability, scatter***, or ***spread***) is the extent to which a [distribution](https://en.wikipedia.org/wiki/Probability_distribution) is stretched or squeezed. Common examples of measures of statistical dispersion are the [***variance***](https://en.wikipedia.org/wiki/Variance)***,***[***standard deviation***](https://en.wikipedia.org/wiki/Standard_deviation), and ***[interquartile range](https://en.wikipedia.org/wiki/Interquartile_range" \o "Interquartile range)***.

**Standard Deviation**:-

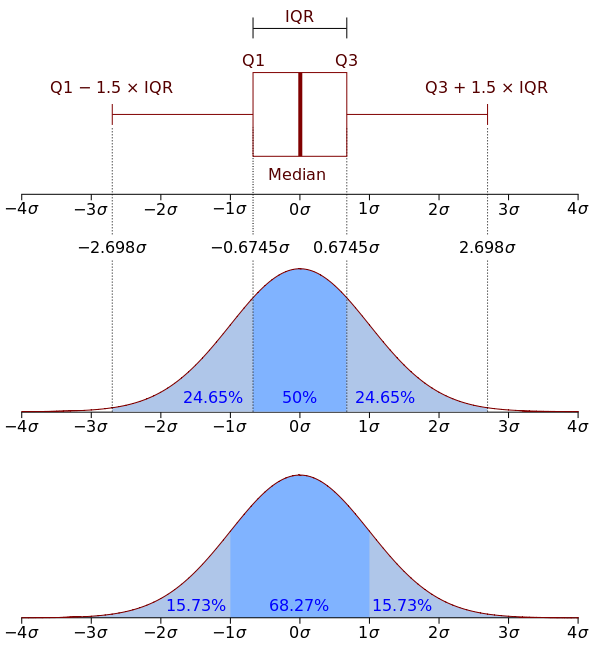
In [statistics](https://en.wikipedia.org/wiki/Statistics), the ***standard deviation*** (SD, also represented by the Greek letter sigma [***σ***](https://en.wikipedia.org/wiki/Sigma) or the Latin letter [***s***](https://en.wikipedia.org/wiki/S)) is a measure that is used to quantify the amount of variation or [dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion) of a set of data values. A low standard deviation indicates that the data points tend to be close to the [mean](https://en.wikipedia.org/wiki/Mean) (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. The ***standard deviation*** of a [random variable](https://en.wikipedia.org/wiki/Random_variable), [statistical population](https://en.wikipedia.org/wiki/Statistical_population), [data set](https://en.wikipedia.org/wiki/Data_set), or [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) is the [***square root***](https://en.wikipedia.org/wiki/Square_root)***of its***[***variance***](https://en.wikipedia.org/wiki/Variance).

**Interquartile range**:-

In [descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics), the interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of [statistical dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion), being equal to the difference between 75th and 25th [percentiles](https://en.wikipedia.org/wiki/Percentiles), or between upper and lower [quartiles](https://en.wikipedia.org/wiki/Quartile),

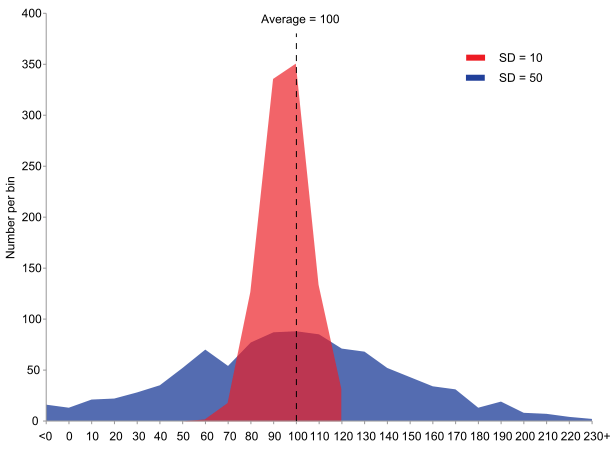
IQR = Q3 −  Q1.

In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a [box plot](https://en.wikipedia.org/wiki/Box_plot) on the data. It is a [trimmed estimator](https://en.wikipedia.org/wiki/Trimmed_estimator), defined as the 25% trimmed [range](https://en.wikipedia.org/wiki/Range_(statistics)), and is the most significant basic [robust measure of scale](https://en.wikipedia.org/wiki/Robust_measures_of_scale).



**Variance**:-

In [**probability theory**](https://en.wikipedia.org/wiki/Probability_theory)**and**[**statistics**](https://en.wikipedia.org/wiki/Statistics), variance is the [expectation](https://en.wikipedia.org/wiki/Expected_value) of the squared [deviation](https://en.wikipedia.org/wiki/Deviation_(statistics)) of a [random variable](https://en.wikipedia.org/wiki/Random_variable) from its [mean](https://en.wikipedia.org/wiki/Expected_value). Informally, it measures how far a set of (random) numbers are spread out from their average value. Variance has a central role in statistics, where some ideas that use it include [**descriptive statistics**](https://en.wikipedia.org/wiki/Descriptive_statistics)**,**[**statistical inference**](https://en.wikipedia.org/wiki/Statistical_inference)**,**[**hypothesis testing**](https://en.wikipedia.org/wiki/Hypothesis_testing)**,**[**goodness of fit**](https://en.wikipedia.org/wiki/Goodness_of_fit)**, and**[**Monte Carlo sampling**](https://en.wikipedia.org/wiki/Monte_Carlo_method). Variance is an important tool in the sciences, where statistical analysis of data is common. The variance is the square of the [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation), the second [central moment](https://en.wikipedia.org/wiki/Central_moment) of a distribution, and the [covariance](https://en.wikipedia.org/wiki/Covariance) of the random variable with itself.



Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

Y-Axis= frequency of occurrence

X-Axis = Data value

For symetrical curve, k = mean = mode = median.

Skewness = 0 for this curve

K = mean = median = mode

For right skewed curve, mean will shift towards positive x axis.

This is also called positively skewed curve.

Skewness = +ve for this curve

Mean

Mode Median

For negatively skewed curve, mean shifts towards negative x axis.

Skewness = -ve for this curve

Mean Median Mode

50%

25% 25% Interquartile Range (IQR) = Q3-Q1

Q1 and Q3 are the median for the half of data

Q2 is the median of data.

Q1 Q2 Q3

**Inferential statistic**:-

Probability density function (PDF)

1) Normal Distribution

Mode Value of interest(Salary,age,Score etc)

Characteristics of normal distribution:-

1) It is used for continuous data. For ex:- Throwing a dice will always result in a value in the range 1-6.

2) Area under curve represents probability. Total area under the curve = Total probability = 1.

3) It is a symmetrical curve, so, mean = median = mode.

4) X-Axis is asymptote to normal distribution.

Asymptote is a line to which the line always approaches but never touches.

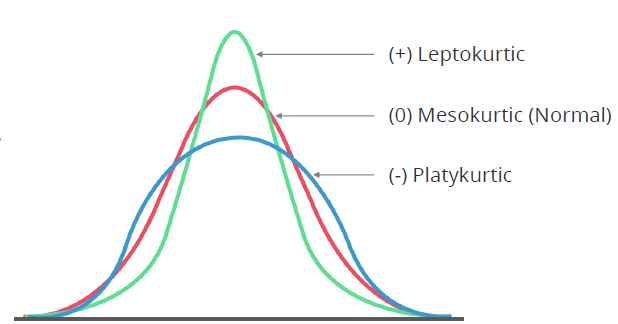
5) Only mean and standard deviation is required to represent normal distribution.

**KURTOSIS:-**

* Kurtosis defines the **shape (TAILEDNESS)** of a probability distribution.
* It measures the tendency of data towards the **centre** or towards the **tail**.

There can be 3 types of curve based on kurtosis:

1. Mesokurtic:- Distributions with zero excess kurtosis are called **mesokurtic**, or **mesokurtotic**(=3). In terms of **shape**, mesokurtic distribution are normally distributed.
2. Leptokurtic:- A distribution with [positive](https://en.wikipedia.org/wiki/Positive_number) excess kurtosis is called **leptokurtic**, or **leptokurtotic**. "Lepto-" means "slender".In terms of **shape**, a leptokurtic distribution has [**fatter tails**](https://en.wikipedia.org/wiki/Fat-tailed_distribution)(>3).
3. Platykurtic:- A distribution with [negative](https://en.wikipedia.org/wiki/Negative_number) excess kurtosis is called **platykurtic**, or **platykurtotic**. "Platy-" means "broad".In terms of **shape**, a platykurtic distribution has **thinner tails**(<3).



**Central Limit Theorem**:-

Central limit theorem (CLT), sets a platform to connect the sample to the population.

1) For a random sample size (**Say10**), for normally distributed population, the distribution of **sample means** is always **normal** but it is **not normal** is **population** is **not normally distributed**.

2) Now, if the sample size is increased (Say 50), the distribution of **sample means** for **normally** and **not normally distributed population is always normal**.

Key Takeaways

* Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.
* Statistical analysis is more reliable when compared to non-statistical analysis.
* Descriptive and inferential are the two major categories of statistics.
* Mean, median, and mode are measures of central tendency, while variance and standard deviation measure the spread of data.
* The spread of distribution is called dispersion and is graphically represented by a histogram and a bell curve.