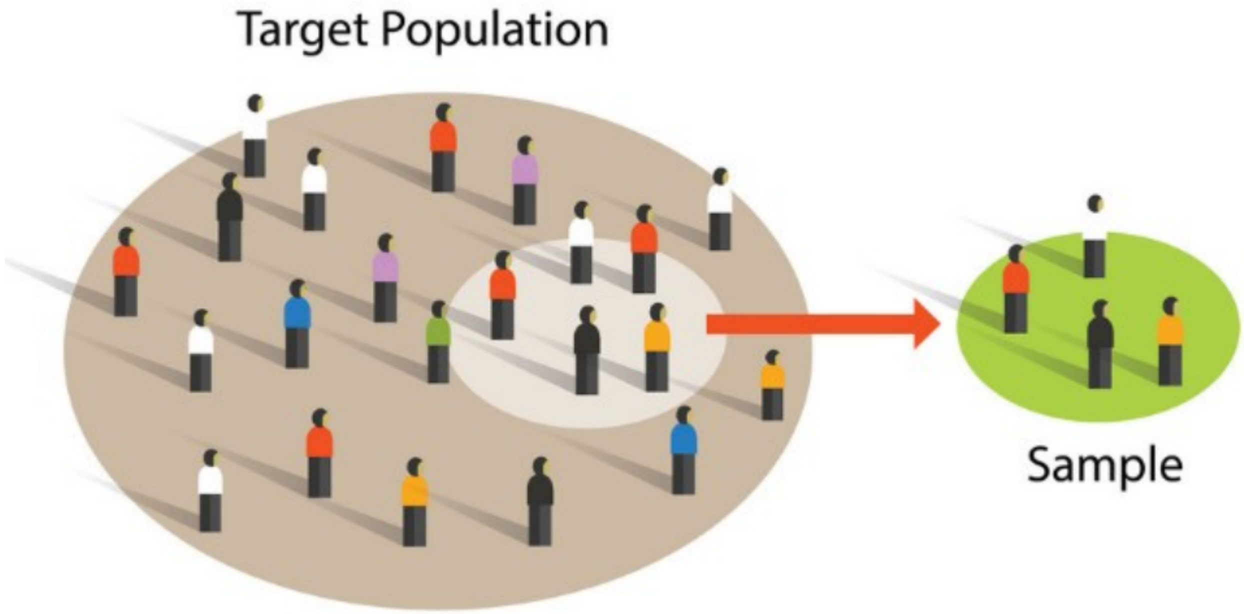
Descriptive statistics summary for Data science

It does exactly as the name suggests 'describe' which summarizes the raw data with help of graphs and overall summary and is easily interpretable by humans. In short it helps us understand "What has happened?"
It contains a summary of definition, formula followed by its advantages and disadvantages , which gives a sense of usage of various statistics in what situation.

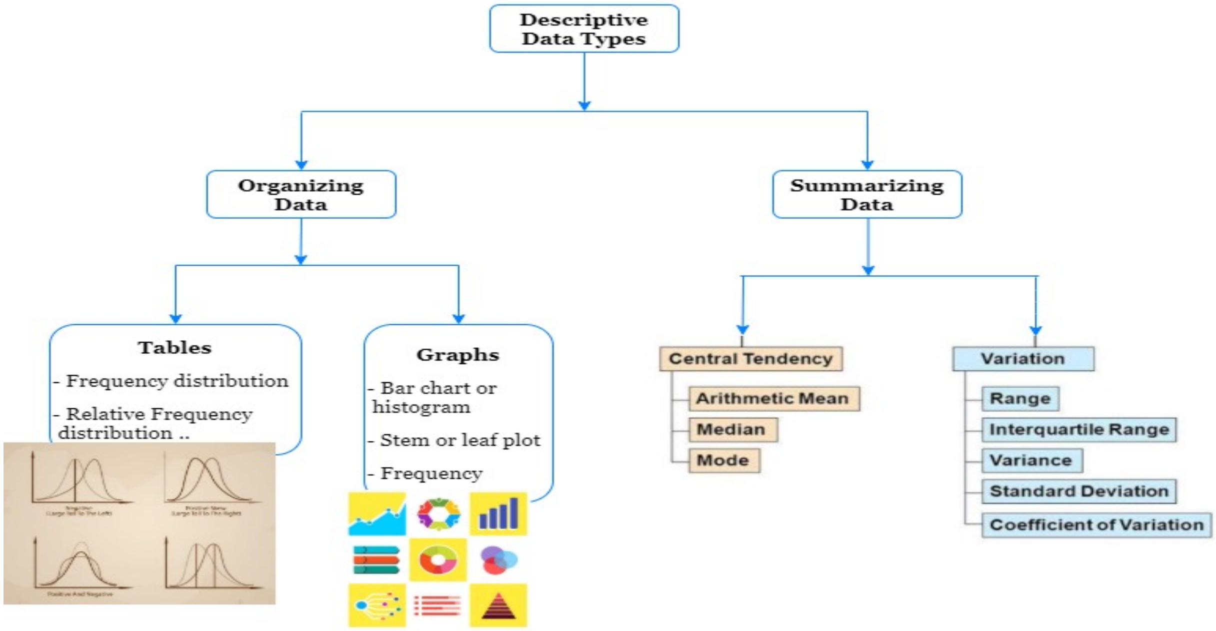
Population vs sample



Population : A data set contain all members of a specified group (the entire list of data

- **Population :** A data set contains all members of a specified group (the entire list of data values).
  Example: The population may be all people living in India.
- **Sample :** A Sample data set contains a part , or a subset of a population. The size of a sample is always less than the size of the population from which it is taken.
  Example: The sample may be some people living in India.

Descriptive data types



Summarizing Data

Measure of Central tendencies

Mean

Most commonly called as average.The mean for a set of data values is the sum of all of the data values divided by the total number of data values.

Formula :

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

Symbolically,

$$\overline{x} = \frac{\sum x}{n}$$

where $\overline{x}$ (read as 'x bar') is the mean of the set of x values,
$\sum x$ is the sum of all the x values, and
n is the number of x values.

Advantages :

- Mean does not require sorting of data, as sorting of data is costly.
- If data is not available at all points, the mode and median will not give correct representation of data.
- It can be used for both continuous and discrete numeric data.

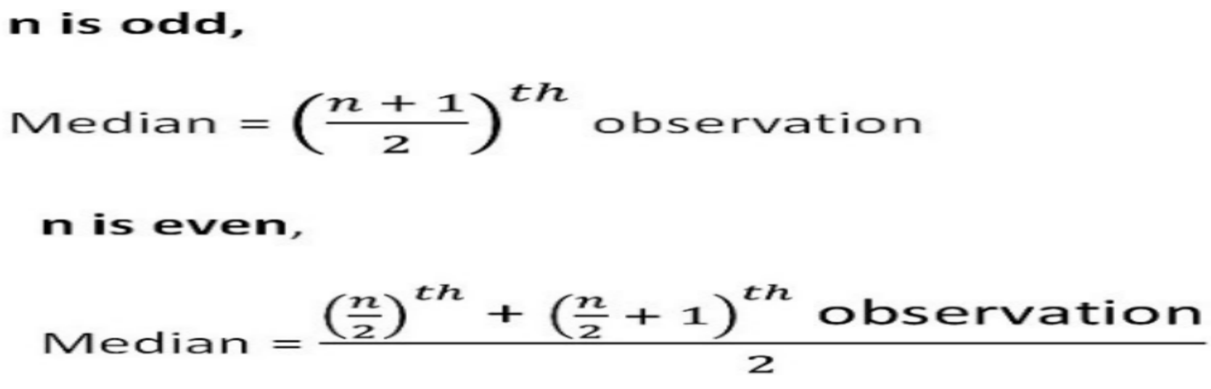
Disadvantages :

- Means can be badly affected by outliers(data point with extreme values unlike the rest).
- The mean cannot be calculated for categorical data, as the values cannot be summed.
- Cannot be graphically inspected/found.

Median

The median of a set of data values is the middle value of the data set when it has been arranged in ascending order. For an odd number of values in the data set the mid number gives the median, while for an even number of values in the data set, average or mean of the mid two values give the median.
When the data are listed in orders, the median is the point at which the 50% of the cases are above and 50% below it is also known as the 50th percentile.

Formula :



n is odd,

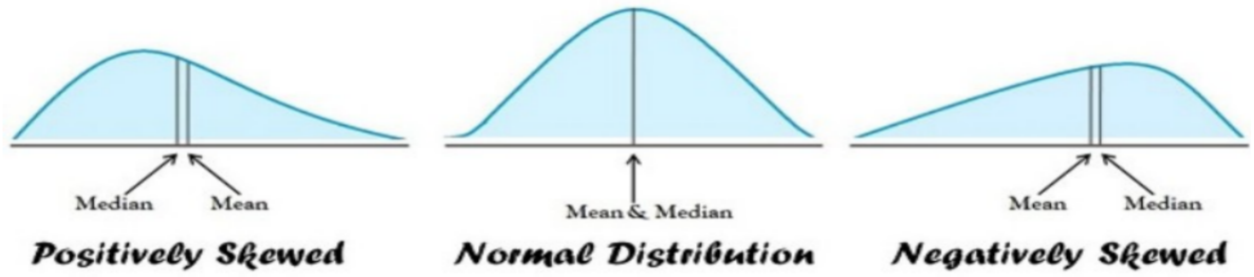
$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

n is even,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

It is unaffected by the outliers and for a symmetric distribution, the mean and median are identical.

In skewed data, the mean lies further towards the skew then the median as shown



Advantages:

- Less affected by outliers and skewed data.
- Can represent data graphically.
- Can be calculated even when No. series is incomplete.

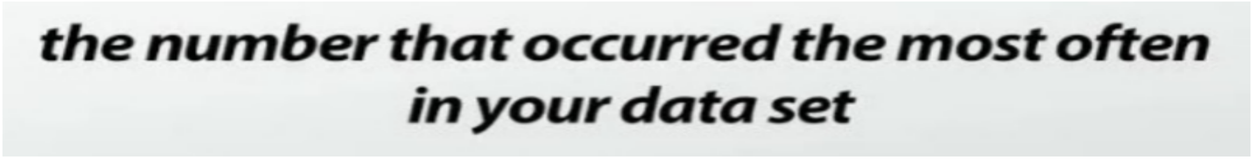
Disadvantages:

- It cannot be identified for the categorical nominal data, as it cannot be logically ordered.
- The sorting of data can be costly sometimes.
- Doesn't account for all the observations.

Mode

Mode is nothing but the most popular number in any given data set or population. It is the value which occurs most frequently in a set of observations. It is possible for the data set to be multimodal (have more than one mode) which means more than one observation has the same number of frequencies.
It may give the most likely experience rather than the "typical" or "central" experience, for example, `` Which size of a shirt should be kept in a store can be decided on the mode value of previous sales of the shirt.

Formula :



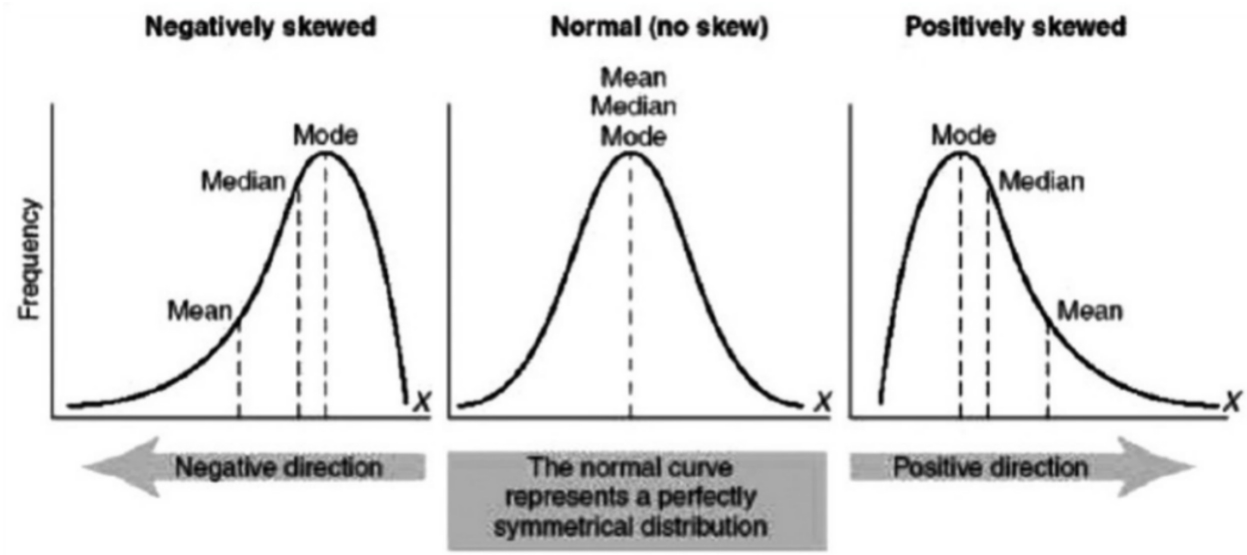
the number that occurred the most often in your data set

Advantages :

- It can be obtained for both numerical and categorical data
- Can be graphically represented with a histogram.

Disadvantages :

- A data set can have one, or more then one , or no mode at all.
- For floating data it will be difficult to calculate the mode
- Could be an inaccurate representation of data as it is not based on all the values.



Measure of Variations

Range

It is the spread or distance between the lowest and highest values of a data set (variables).

Formula :

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Advantages :

- The prime advantage of this measure of dispersion is that it is easy to calculate.

Disadvantages :

- It is very sensitive to outliers and does not use all the observations in a data set.
- It is more informative to provide the minimum and the maximum values rather than providing the range.

Interquartile Range

It is defined as the difference between the (Q1)25th and (Q3)75th percentile (also called the first and third quartile). Hence the interquartile range describes the middle 50% of observations.
If the interquartile range is large it means that the middle 50% of observations are spaced wide apart.

Formula :

$$\mathrm{IQR} = Q_3 - Q_1$$

Advantages :

- It can be used as a measure of variability if the extreme values are not being recorded exactly (as in case of open-ended class intervals in the frequency distribution).
- It is not affected by extreme values.

Disadvantages :

- The main disadvantage in using interquartile range as a measure of dispersion is that it is not amenable to mathematical manipulation.

Variance

Variance ($\sigma2$) in statistics is a measurement of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set.

Formula :

$$\text{variance } \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

where:

x_i = the i^{th} data point

$\bar{x}$ = the mean of all data points

n = the number of data points

Statisticians use variance to see how individual numbers relate to each other within a data set, rather than using broader mathematical techniques such as arranging numbers into quartiles.

Advantages :

- The advantage of variance is that it treats all deviations from the mean the same regardless of their direction. The squared deviations cannot sum to zero and give the appearance of no variability at all in the data.

Disadvantages :

- It gives added weight to outliers, the numbers that are far from the mean. Squaring these numbers can skew the data.
- It is not easily interpreted as we square the data, changing its dimensions from the original one.

Standard deviation(SD)

The problem with variance is that it cannot give the correct representation of the deviation as the result is squared and is in a different unit from the normal set. To overcome this problem we calculate the SD

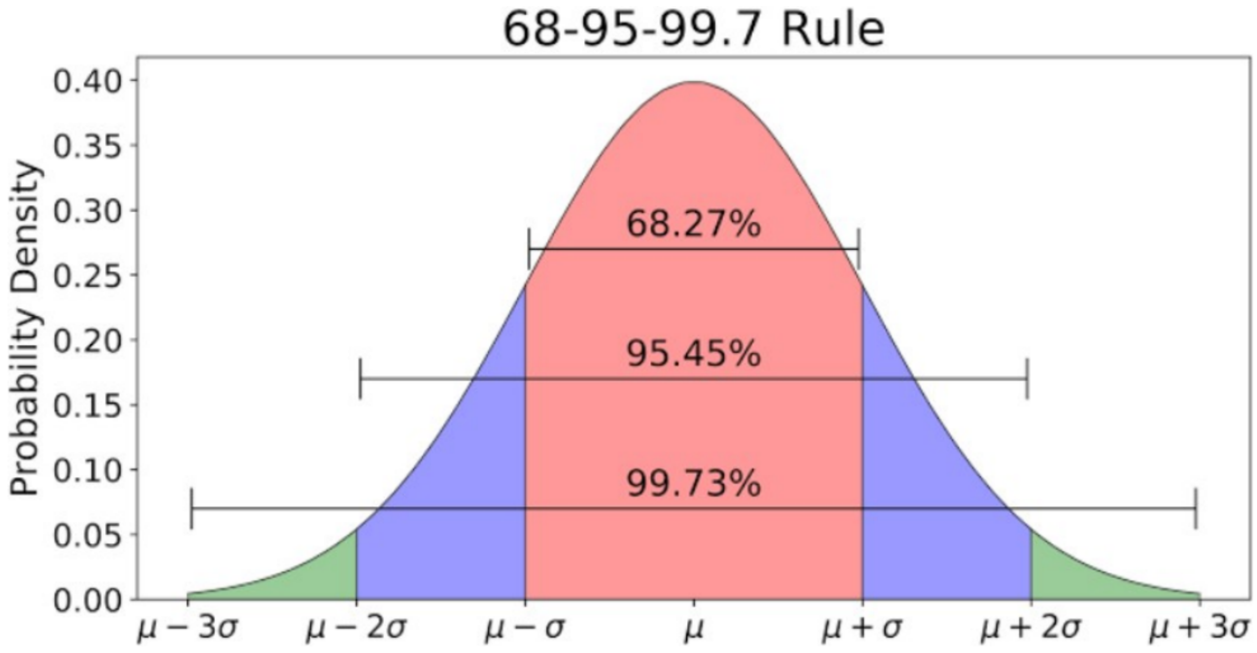
Standard deviation (SD) is the most commonly used measure of dispersion. It is a measure of spread of data about the mean. SD is the square root of the sum of squared deviations from the mean divided by the number of observations.

Formula :

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{n}}$$

where,

σ = population standard deviation
$\sum$ = sum of...
μ = population mean
n = number of scores in sample.



Advantages :

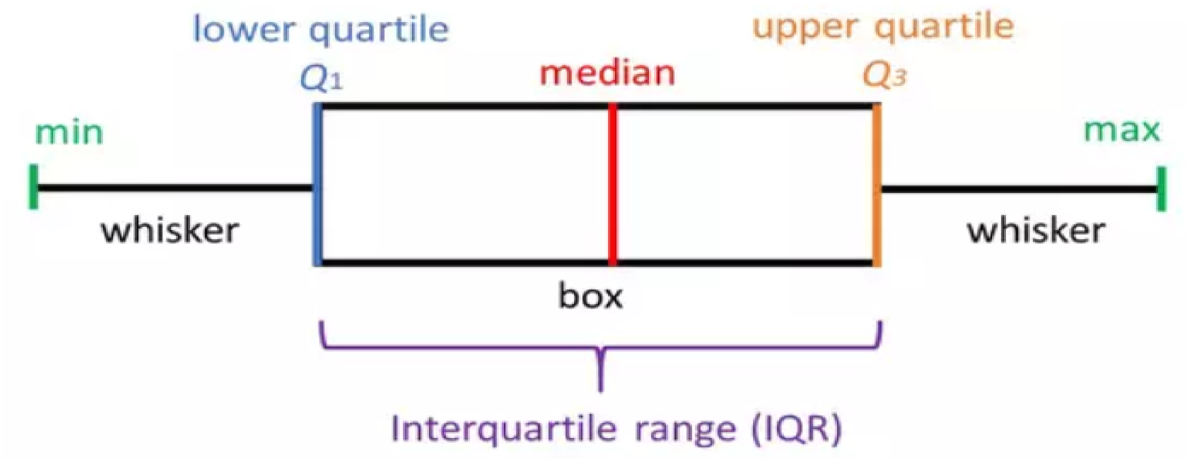
- The reason why SD is a very useful measure of dispersion is that, if the observations are from a normal distribution, then 68% of observations lie between mean ± 1 SD 95% of observations lie between mean ± 2 SD and 99.7% of observations lie between mean ± 3 SD
- The other advantage of SD is that along with means it can be used to detect skewness.

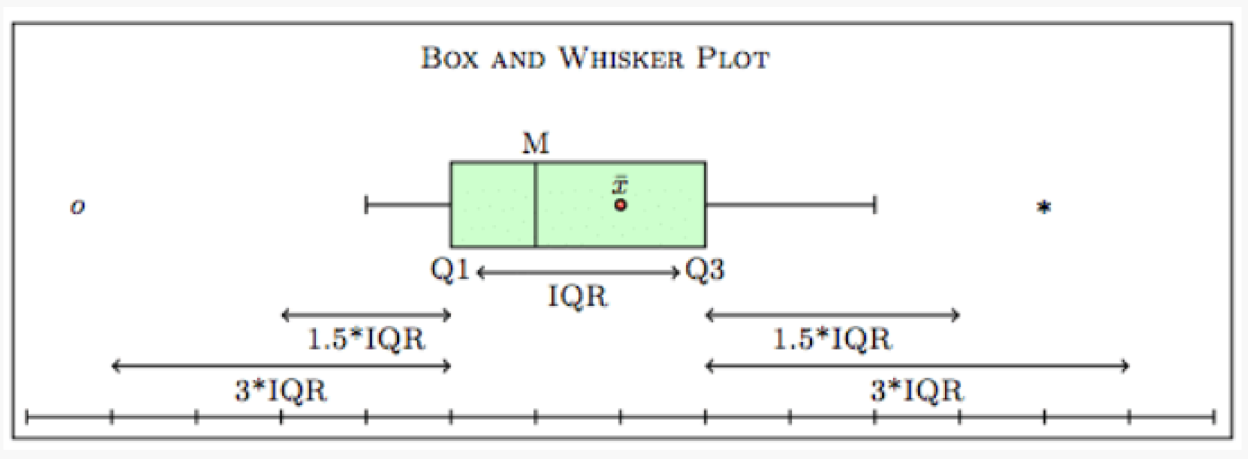
Disadvantages :

- It is an inappropriate measure of dispersion for skewed data.

Box plot

Box plot help us depict the descriptive statistics data graphically.


lower quartile
Q1
median
upper quartile
Q3
min
max
whisker
whisker
box
Interquartile range (IQR)


Box and Whisker Plot
M
$\bar{x}$
o
*
Q1
Q3
IQR
1.5*IQR
1.5*IQR
3*IQR
3*IQR