STATISTICS FOR DATA SCIENCE

Statistics is the science that study about the collection, analysis, interpretation, presentation, and organization of data

There are two types of statistics
A. Descriptive Statistics
B. Inferential Statistics

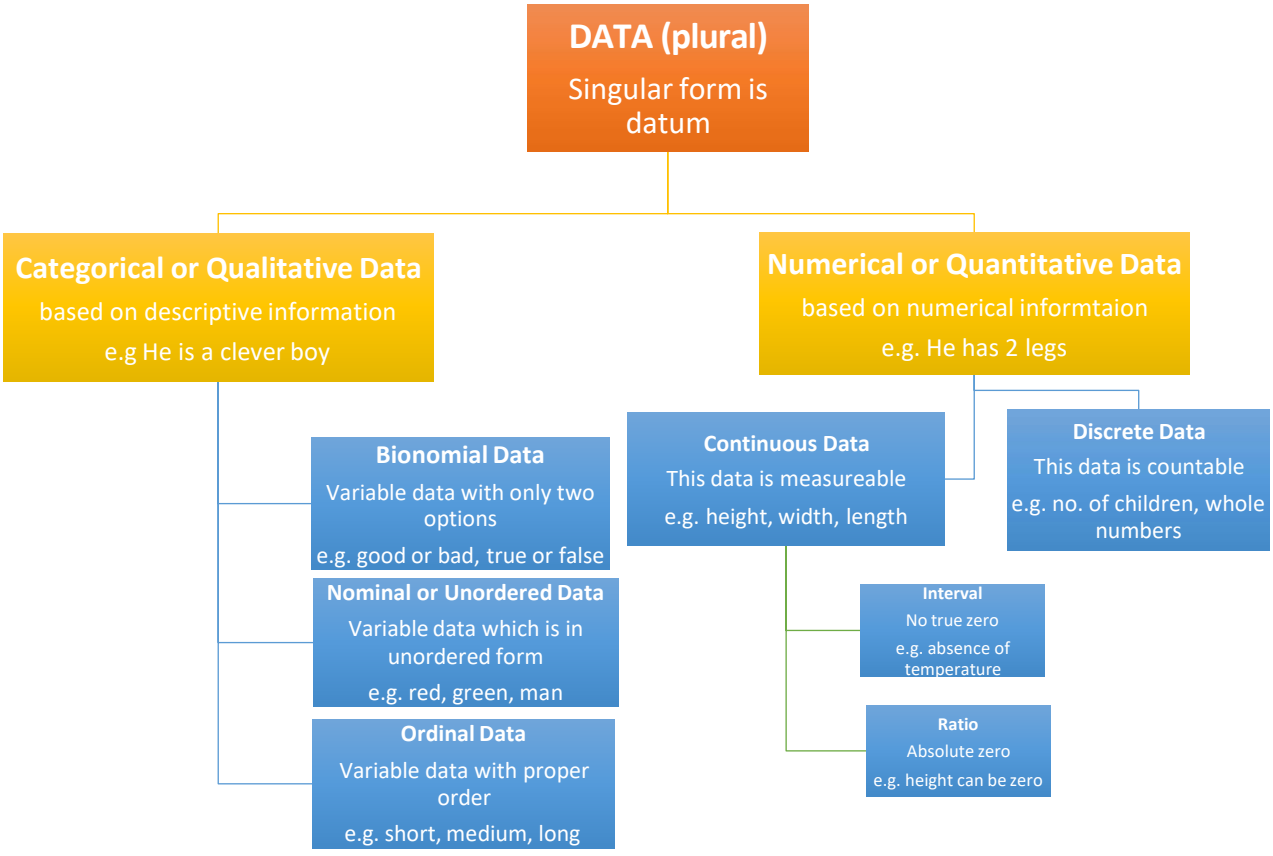
A. DESCRIPTIVE STATISTICS:

Before going to discuss about descriptive statistics, first we recall the basic concept of data and its types again here before starting descriptive statistics …..

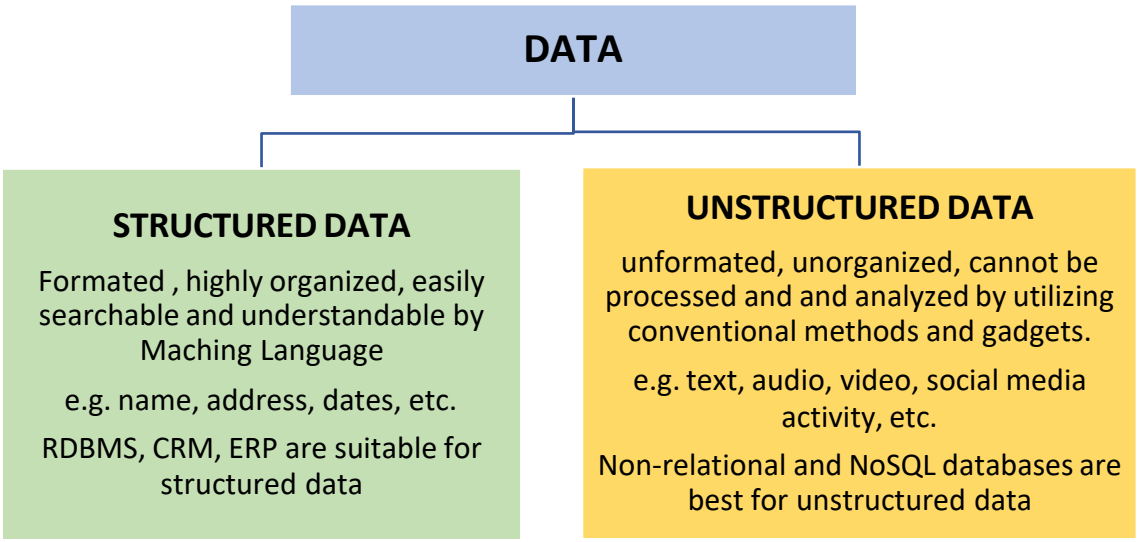
Data:

Data is a collection of factual information based on numbers, words, observations, measurements which can be utilized for calculation, discussion and reasoning.

TYPES OF DATA:



The crude dataset is the basic foundation of data science and it may be of different kinds like Structured Data (Tabular structure), Unstructured Data (pictures, recordings, messages, PDF documents and so forth.) and Semi Structured.



Furthermore, there are two kinds of data i.e. population data and sample data.

- **Population Data:**

Population data is the collection of all items of interest which is denoted by 'N' and the numbers we obtained when using population are called parameters.

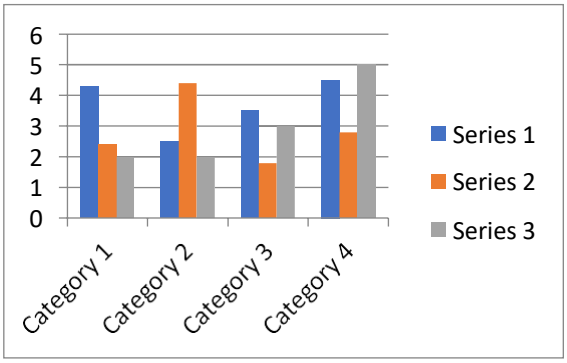
- **Sample Data:**

Sample data is a subset of the population which is denoted by 'n' and the numbers we obtained when using sample are called statistics.

Graphical Representation of variables in form of Graph & Tables:

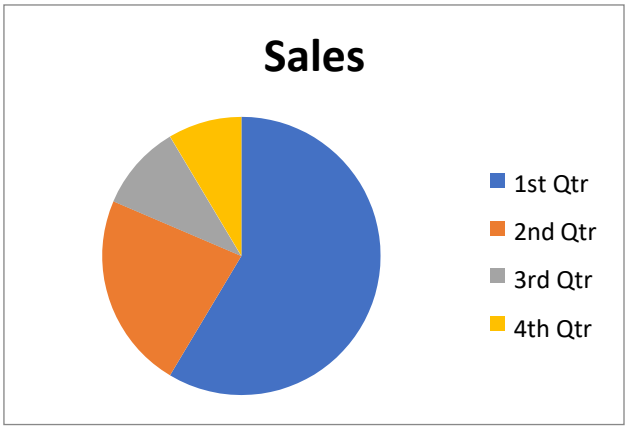
i. Bar Chart:

Bar charts are frequently being used to display data. In bar chart, each bar represents a category and y-axis shows the frequency as shown in figure



ii. Pie chart:

Pie Charts are frequently being used to display market share. If we want to see the share of any item as a part of the total then we utilized pie chart, as shown in figure below:



iii. Frequency Distribution Table:

Frequency distribution table shows the category and its corresponding absolute frequency as shown in figure

Category	Frequency
Black	12
Brown	5
Blond	3
Red	7

Relative frequency = frequency / total frequency

Measures of central tendency:

It is a single value that explains a set of data by identifying the central positing within that set of data. Measure of central tendency is also called measure of central location. The measures of central tendency are:

i. Mean
ii. Median
iii. Mode

i. Mean:

It is most popular to measures the central tendency. It is used with both discrete and continuous data. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. Therefore, if we have n values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by $\bar{x}$ is given by,

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

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

If we intend to calculate the population means instead of sample mean then we use the greet letter μ as

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

ii. Median:

It is the mid score of a dataset that has been arranged in order of magnitude. In order to calculate the median, suppose we have the following dataset:

10	20	30	15	20	30	15	20	30	15	20

First of all, we re-arrange this data into order of magnitude from smaller to larger

10	15	15	15	20	**20**	20	20	30	30	30

Therefore, in this case bold figure 20 is our median. It is the middle mark, as there are 5 scores before it and 5 scores after it. However, if we have an odd number of scores like this one,

10	20	30	15	20	15	20	30	15	20

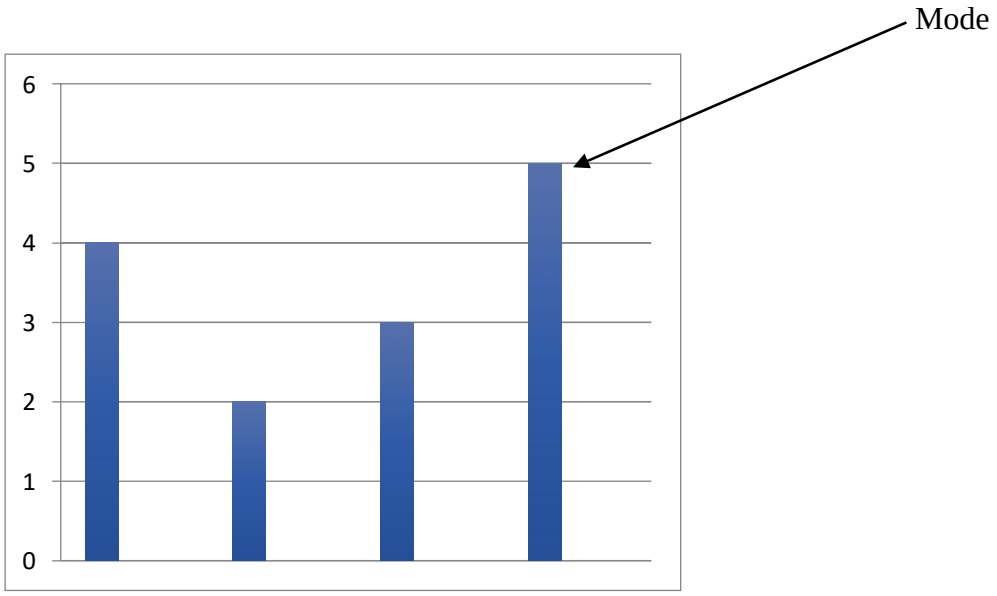
Then, we re-arrange this data into order of magnitude and we obtain

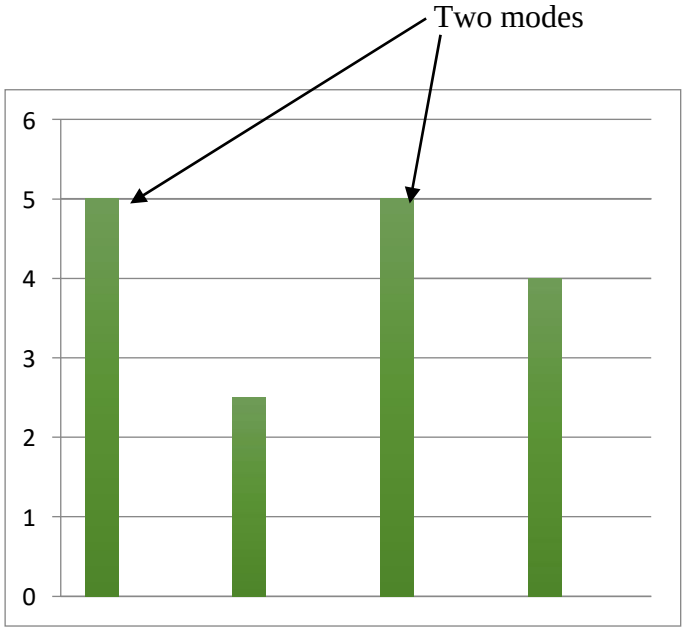
10	15	15	15	**20**	**20**	20	20	30	30

In this case, we have to take two values i.e. 20, 20 and average them to get a median i.e. 20.

iii. Mode:

It is a value that most often score in our dataset. A dataset can have no mode, one mode or multiple modes. It can be calculated by finding the value with the maximum frequency. For instance,





Measure of Asymmetry:

Skewness:

It is the measure of asymmetry that shows whether the observations in a dataset are focused on one side. Skewness can be calculated by the following formula

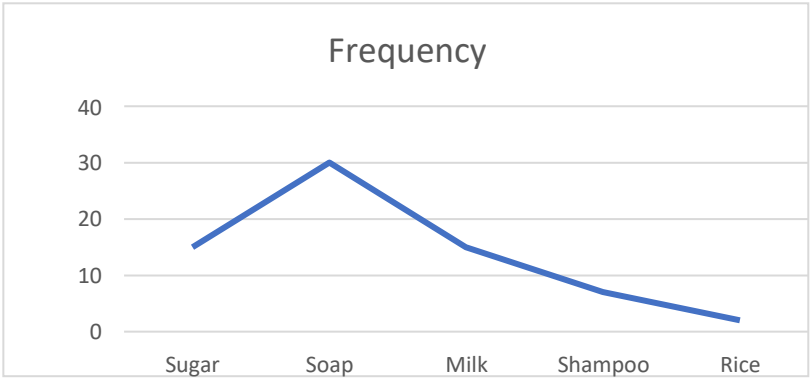
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i-}\bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i-}\bar{x})^2}^3}$$

There are two types of skewness,

i. Right or Positive Skewness
ii. Left or Negative Skewness

i. Right or Positive Skewness:

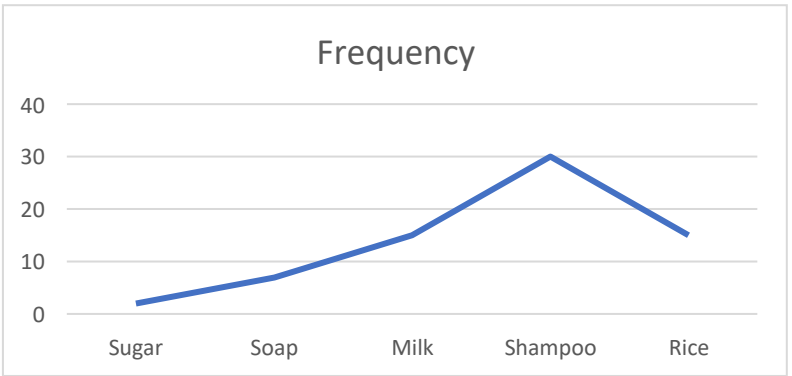
Right or positive skewness means that the outliers are to the right (long tail to the right) as shown in figure



Mean > median>Mode

ii. Left or Negative Skewness:

Left or negative skewness means that the outliers are to the left (long tail to the left) as shown in figure



Mean < median< mode

However, if mean = median = mode then no skew and therefore, distribution will be symmetrical.

Variance and Standard Deviation:

Variance and Standard Deviation measure the dispersion of a set of data points around its means value.

Sample Variance formula:

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

Population Variance formula:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Sample Standard Deviation formula:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Population Standard Deviation formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Coefficient of Variation:

There is no unit of measurement for Coefficient of variation. Coefficient of variation is perfect for comparison and universal across datasets. Formula of coefficient of variation is given below: -

$$CV = \frac{S}{\bar{x}}$$

Covariance and correlation:

Covariance	Correlation
Covariance is a statistical measure which is defined as a systematic relationship between a pair of random variables wherein change in one variable responded by an equivalent change in another variable	Correlation is a statistical measure which is defined as a systematic relationship between a pair of random variables wherein movement in one variable responded by an equivalent movement in another variable
The value of covariance lies between -∞ and +∞	The value of correlation lies between -1 and +1
A covariance of 0 means that the two variables	A correlation of 0 means that the two variables

are independent.	are independent
A positive covariance means that two variables move together	A correlation of 1 means perfect positive correlation
A negative covariance means that the two variables move in opposite directions	A correlation of -1 means perfect negative correlation.
Sample Covariance formula: $S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$ Population Covariance formula: $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$	Sample Correlation formula: $r = \frac{s_{xy}}{s_x s_y}$ Population Correlation formula: $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

B. INFERENTIAL STATISTICS:

i. Normal Distribution:

It is also known as Gaussian distribution or Bell Curve. It is mostly used in regression analysis. A lot of things closely follow this distribution:

- heights of people

- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test
- stock market information

When data is normal distributed then distribution is symmetric and

Mean = median = mode

$$N \sim(\mu, \sigma^2)$$

Where, N for normal, ~ for distribution, μ is mean, and σ^2 is the variance

ii. Standard Normal Distribution:

It is a normal distribution with a mean of 0 and a standard deviation of 1. Every normal distribution can be standardized using the following formula

$$a = \frac{x - \mu}{\sigma}$$

$$N \sim(0, 1)$$

Standardization permit to compare different normally distributed datasets, test hypothesis, detect outliers and normality, create confidence intervals and perform regression analysis.

Quartile

Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q1, Q2, and Q3, respectively. Q2 nothing but the median, since it indicates the position of the item in the list and thus, is a positional average.

Quartile formula

Q1 = [(n+1)/4]th item

Q2 = [(n+1)/2]th item

Q3 = [3(n+1)/4]th item

- First quartile: 25% from smallest to largest of numbers
- Second quartile: between 25.1% and 50% (till median)
- Third quartile 51% to 75% (above the median)
- Fourth quartile: 25% of largest numbers

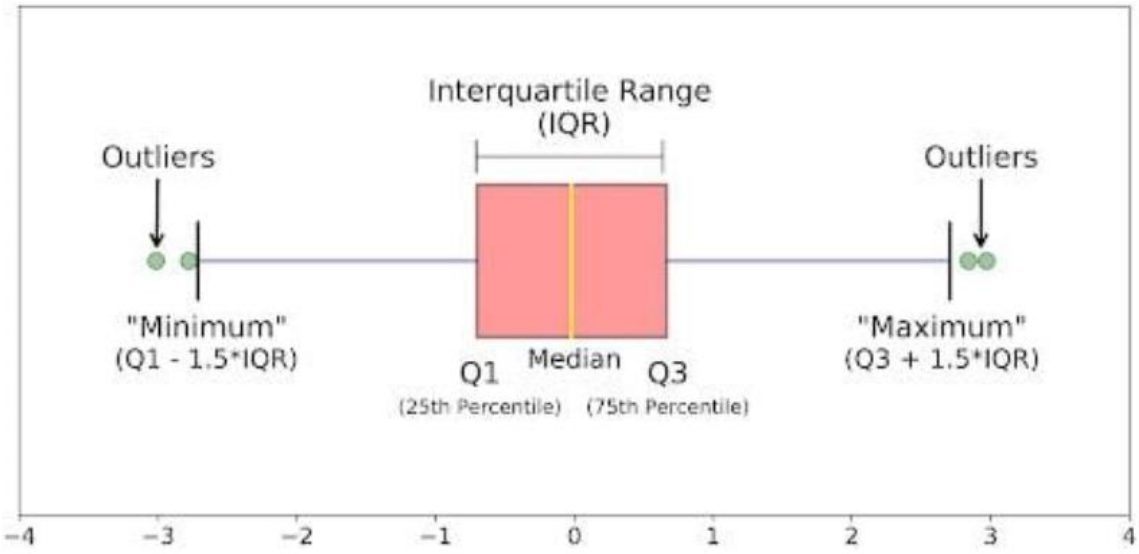
Interquartile range:

Interquartile range is the difference between the upper and lower quartile of a given data set and is also called a midspread. It is measure of statistical distribution, which is equal to the difference between the upper and lower quartiles.

IQR = Q3-Q1

Box plot

A boxplot also known as box and whisker plot is a standardized way of displaying the distribution of data set based on its five-number summary of datapoint: the ‘minimum’, ‘first quartile’ [Q!], median, third quartile[Q3] and ‘maximum’,



Outliers

Outliers is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set.

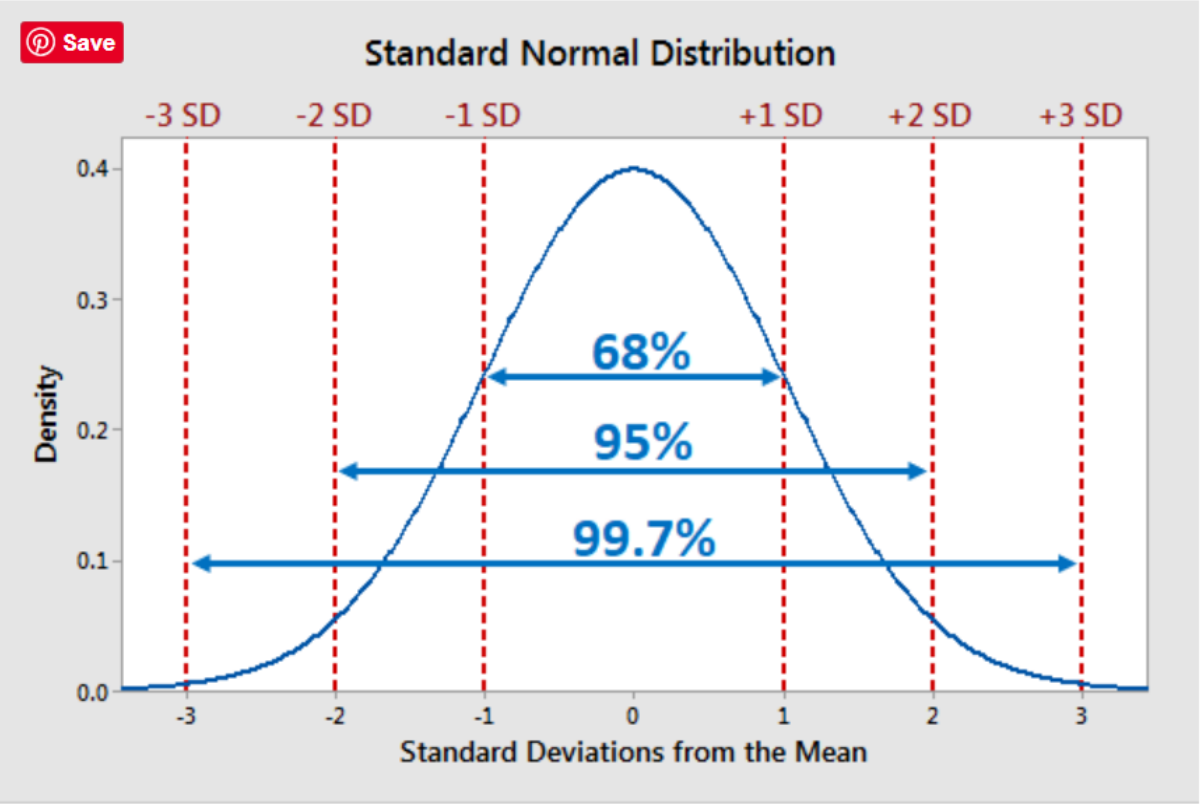
Outlier may occur due to the variability in the data, or due to experimental error/human error.

Deal with outliers

- Drop the outliers
- Fiil with Median value
- Cap the value with Q_3 and Q_1

Empirical formula

The empirical rule in statistics, also known as the 68 95 99 rules, states that for normal distributions, 68% of observed data points will lie inside one standard deviation of the mean, 95% will fall within two standard deviations, and 99.7% will occur within three standard deviations.



Chebyshev's Inequality

Chebyshev's inequality is a probability theory that guarantees that within a specified range of distance form the mean, for a large range of probability distributions, no more than a specific fraction of values will be present.

The fraction for which no more than a certain number of values can exceed is represented by $1/k^2$

Chebyshev's inequality states that within two standard deviations away from the mean contains 75% of the values, and within three standard deviations away from the mean contains 88.9% of the values. It holds for a wide range of probability distributions, not only normal distribution.

Z- standardization(z-score)

Z-score is a statistical measurement that describes a value's relationship to the mean of a group of values. Z-values is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point score is identical to the mean score. A z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-score may be positive or negative, with positive value indicating the score is above the mean and negative score indicating it is below the mean.

Z-score Formula

The statistical formula for value's z-score is calculated using the following formula:

$Z = (x-\mu)/\sigma$

Z = Z-score

X = the value being evaluated

μ = the mean

σ = the standard deviation

Normalisation:

In statistics, the term 'normalization" refers to the scaling down the data set such that the normalized data falls between 0 and 1. This normalization technique helps compare corresponding normalized values from two or more dataset.

Normalization Formula

$$X \text{ normalized} = \frac{(X - X \text{ minimum})}{(X \text{ maximum} - X \text{ minimum})}$$