**Statistics and Probability**

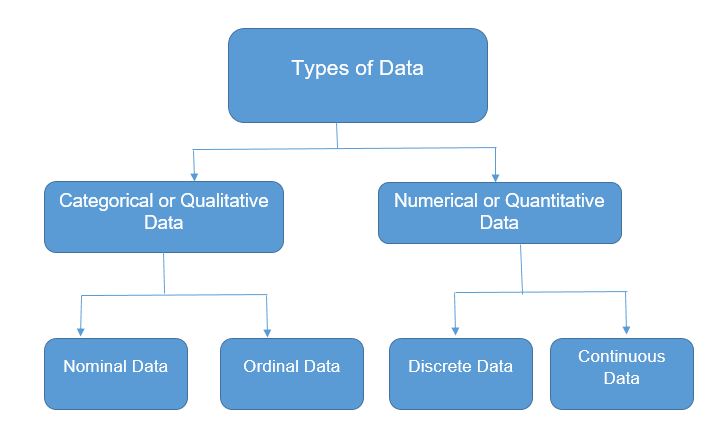
**Data**:

Data refers to facts and statistics collected together for reference and analysis.

|  |  |
| --- | --- |
| Data can be | Collected and stored |
| Measured |
| Analysed |
| Visualized |

**Categories of data:**

Data is of 2 types:



**Qualitative Data:**

Data deals with characteristics and descriptors that can’t easily measure, but can be observed subjectively.

Ex: Name, Gender (Male/Female), Type of coins in a jar

The qualitative data is again divided into 2 types:

* **Nominal Data**: Data with no inherent order or ranking such as gender or race, such kind of data is called nominal data.

**Ex**: country, gender, race, hair colour etc. of a group of people.

* **Ordinal Data:** Data with an ordered series is called ordinal data type.

**Ex:**

|  |  |
| --- | --- |
| **Customer ID** | **Rating** |
| 001 | Good |
| 002 | Bad |
| 003 | Average |

**Quantitative Data:**

Data with numbers and things that can be measured objectively.

Ex: 1) The apartment is 550 square feet in size.

2) 98% of the graduating high school class is going to college

The quantitative data is again divided into 2 types:

* **Discrete Data:** Also known as categorical data, it can hold finite number of possible values.

**Ex:** Number of students in a class.

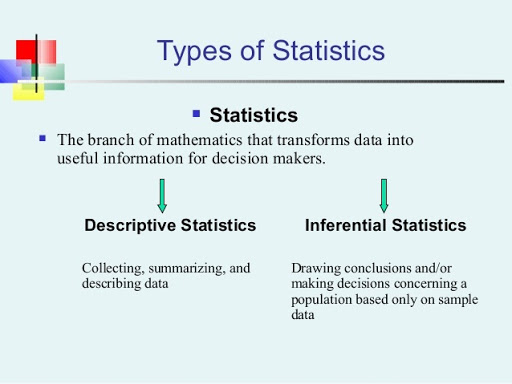
* **Continuous Data:** Data that can hold infinite number of possible values.

**Ex:** Weight of a person.

**Statistics:**

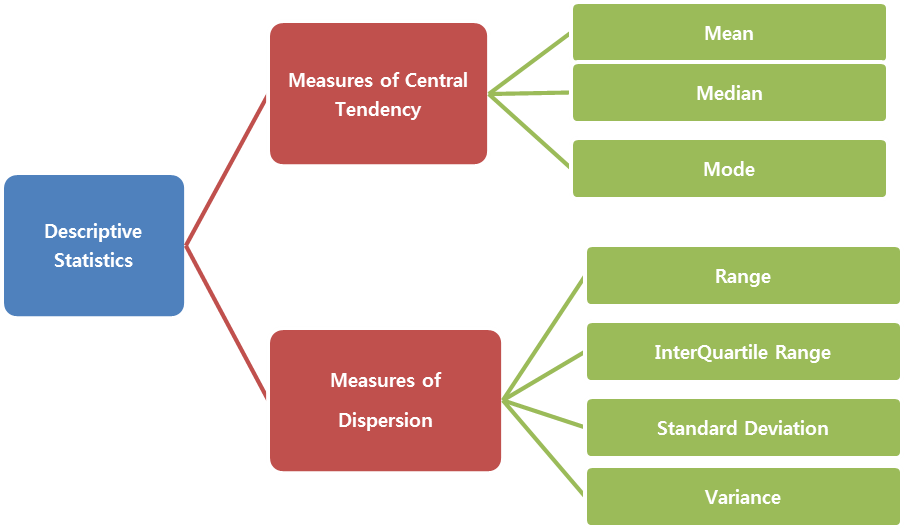
Statistics is an area of applied mathematics concerned with the data collection, organizing, presenting, analysing and interpreting data to make effective decision.

Implemented to manipulate, summarize and investigate data for decision making.



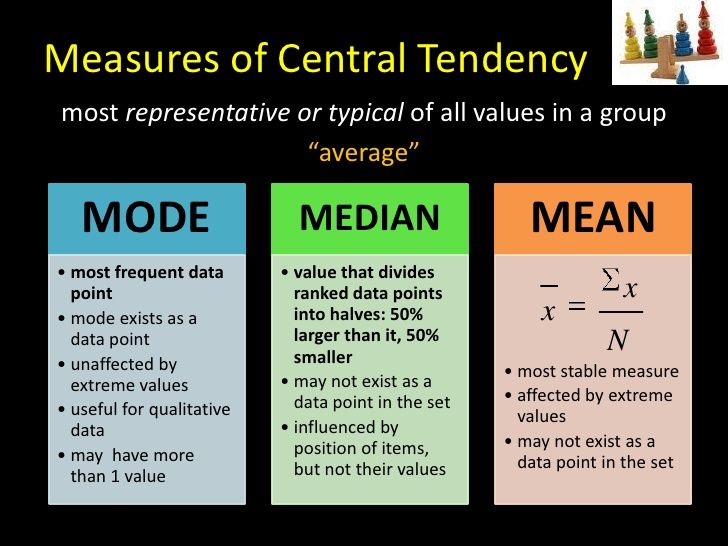
**Descriptive Statistics:**

* Descriptive statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.
* Descriptive statistics is mainly focused upon main characteristics of the data. It provides graphical summary of the data.
* There are two categories of descriptive statistics.



**Measures of Central Tendency:**

A measure of central tendency is a summary statistic that represents the centre point or typical value of a dataset. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.



**Problems on Measures of Central Tendency:**

1. Find the mean, median and mode of the following figures.

30 , 41 , 47 , 54 , 23 ,34 ,37 , 51 , 53 , 47

1. Find the Mean, Median and mode of the following figures.

17 , 19 , 21 , 13, 17 , 16 , 17 , 13 , 18 , 24 , 22 , 20

1. Find the mode from the following data:

12 , 14 , 16 , 18 , 26 , 16 , 20 , 16 , 11 , 12 , 16 , 16 , 20 , 24

**Measures of Dispersion/Spread:**

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range.

**Range:**

Range is the given measure of how spread apart the values in a dataset are.

**Range = Max Value – Min Value**

**Coefficient of Range = Max – Min / Max +Min**

**Ex:** Find the range and coefficient of range of the following data:

3 , 7 , 21 , 24 , 37 , 40 , 45

**Ans**: Range = Max Value – Min Value

= 45-3

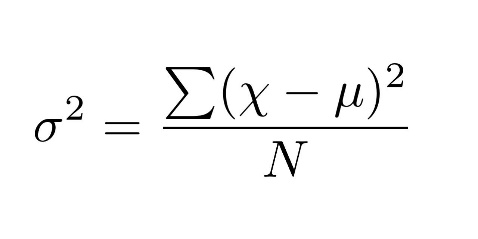
= 42

Coefficient of Range = Max – Min / Max + Min

= 45-3/45+3 = 0.875

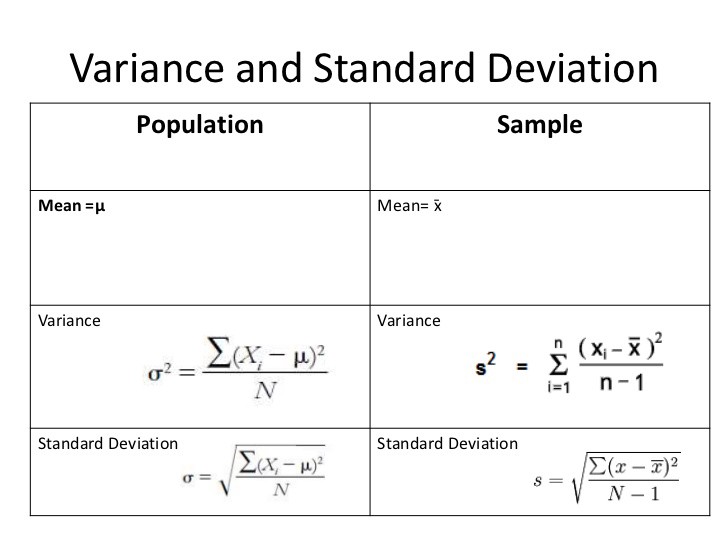
**Variance:**

Variance describe how much a random variable differ from its expected value. It entails computing squares of deviation.



**Standard Deviation:**

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.



**Problems on Variance and Standard Deviation:**

1. Calculate the Variance and standard deviation of the following

25 , 27 , 31 , 32 , 35

1. Find the Range and co efficient of range

14 , 18 ,19 , 21 ,23 ,25 , 27 , 29 , 33 , 34 , 37

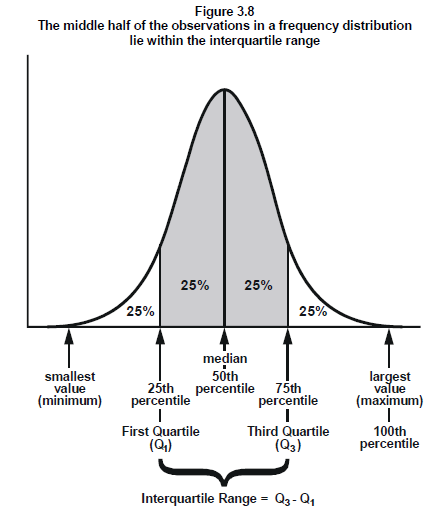
**Measures of Quartiles:**

**Quartiles:**

Quartiles tell us about the spread of a data set by breaking the data set into quartiles, just like the median breaks it in half.

**Inter Quartile Range:**

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

****

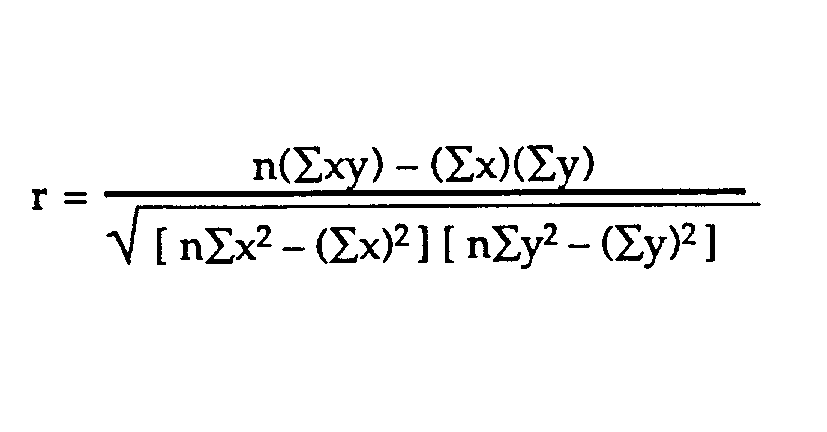
* Outlier is who performance will far from the average performance. The outlier performance will affect the **mean** a lot.
* Boxplot are used to for data visualization to check the presence of outliers.
* High value outliers > Q3
* Low value outlier < Q1
* To calculate outlier, we need to calculate quartiles.

**Bi-variate / Correlation and Covariance:**

**Correlation:**

* Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related.
* When both the variables are moving the same direction, it is called positive correlation.
* When both the variables are inversely related then it is called as negative correlation.

**Pearson’s Coefficient Correlation:**

Karl Pearson’s Coefficient of Correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linear related variables. The coefficient of correlation is expressed by **“r”.**

* The [correlation coefficient lies between -1 and +1](http://itfeature.com/correlation-and-regression-analysis/correlation-coefficient-lies-between-1-and-1), symbolically −1≤r≤1
* If r = 0, there is no correlation.
* r = 1 Perfect positive correlation
* r= 0 <=0.5 – Low positive correlation
* r=0.5 to 1 – High positive correlation
* r= -1 Perfect negative correlation
* r= 0 <=-0.5 – Low negative correlation