

***Module 2***

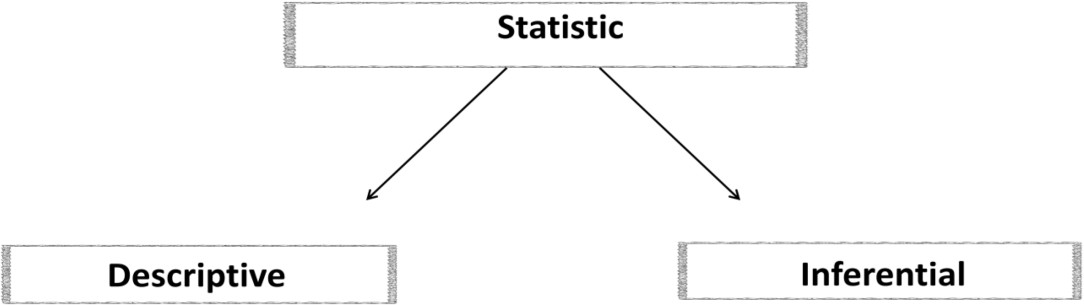
Statistical Fundamental-I

## What is Statistics-

Statistics is a branch of Mathematics dealing with Data Collection, Organization, Analysis, Interpretation and Presentation.

As defined by the American Statistical Association (ASA)- is the science of learning from data and of measuring, controlling and communication uncertainty.

## Types of Statistics



1. **Descriptive Statistics**

It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

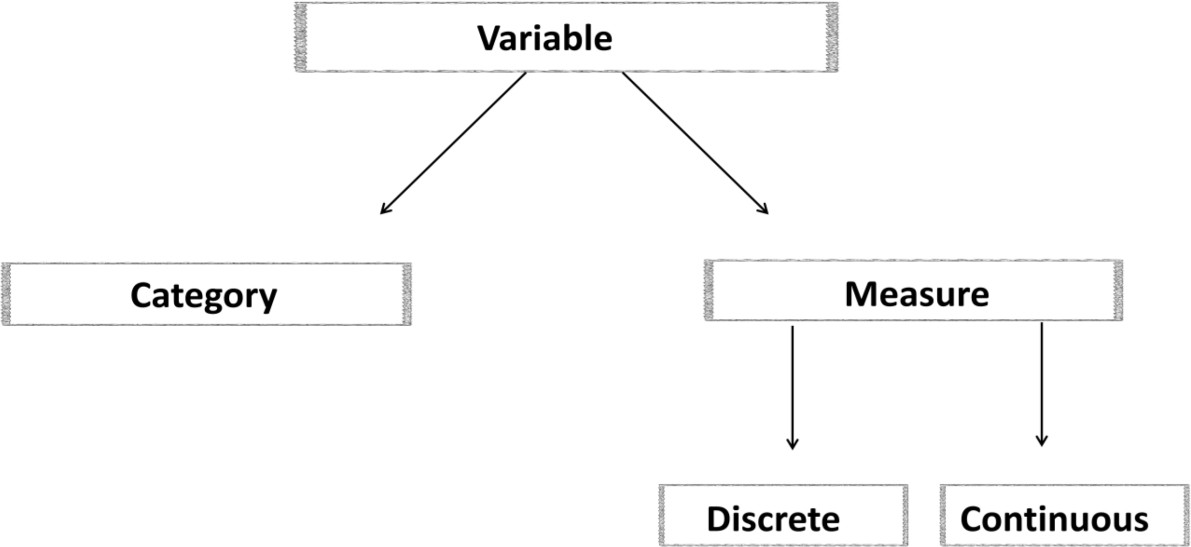
### Inferential Statistics

Inferential Statistics is used to draw inferences beyond the immediate data available. We can answer the following questions with the help of inferential statistics: making inferences about the population from the sample.

|  |  |
| --- | --- |
| **Descriptive Statistics** | **Inferential Statistics** |
| Descriptive statistics work with smaller  data. There is no need for sampling and the entire population data is available. | Inferential statistics work with large data  set. Analyzing entire population based on sample parameter is a strength. |
| Process is simpler to do | Process is complex as we have to decide best sampling technique. |
| Descriptive statistics are likely to be 100% accurate because there is no assumption. | This is not 100% accurate. inferential  statistics always make inference about larger population based on sample. |
| Find results are shown in form of charts, tables and graphs. | Find result in probability score. |
| Tool-Measure of central tendency (mean, median, mode) spread of data (range  standard deviation) | Tool-hypothesis test, analysis of variance. |
| Organize, analyze and present data in meaningful manner. | Compress, test and predicts future outcome. |

## Basic Terminologies

1. **Population:** It is the set of sources from which data has to be collected. Population in statistics include all members of a define group that we are studying or collecting information on for data driven decisions.
2. **Sample:** It is a subset of the population
3. **Variable**

* A variable is named value that changes
* Simply something that varies like
* Your weight
* Environment Temperature
* Your Marks

1. **Measure Variable:** It has values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore, numeric variables are quantitative variables.

* **Discrete Variable:** It is a numeric variable. Observations can take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value.

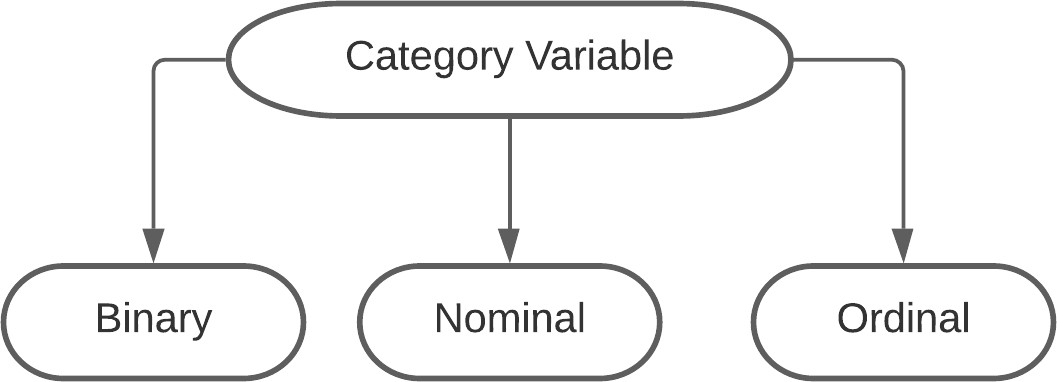
**Examples:**

* + The number of registered cars, number of business locations
  + And number of children in a family
  + All of which measured as whole units (i.e. 1, 2, 3 cars).
* **Continuous Variable:** A continuous variable is defined as a variable which can take an uncountable set of values or infinite set of values.

**Examples:**

* + Height
  + Age
  + Temperature.

1. **Categorical Variable:** It has values that describe a 'quality' or 'characteristic ‘of a data unit, like 'what type' or 'which category'.



|  |  |  |
| --- | --- | --- |
| **Type of Variable** | **Data Represent** | **Example** |
| Binary Variables | Yes/No outcomes | Head/ tails in coin flip  Win/Lose in match |
| Nominal Variables | Group with no rank or order Between them | * Species names * Colors * Brands |
| Ordinal Variables | Group with that are  ranked in a specific order. | Finishing place in a race  rating scale response in a survey |

|  |  |  |
| --- | --- | --- |
| **Type of Variable** | **Data Represent** | **Example** |
| Independed Variables | Variables you manipulate in order to affect the outcome of an experiment. | Sales is independent from Profit. |
| Depended Variables | Variables you manipulate in  order to affect the outcome of an experiment. | Profit depends on Sales. |

***ANALYTICS Versus ANALYSIS***

|  |  |
| --- | --- |
| **Analysis** | **Analytics** |
| We perform analysis on things that  have already happened in past. | Analytics is working on future |
| The why? How? What? Of happened in past. | Analytics is utilizing machine learning, Statistics, algorithm, models to take better decisions and get  better insight from data |
| What we earn last year | Analytics is defined as a process of  Transforming data into action. |

***What is ANALYTICS?***

Analytics provides us with meaningful information which may otherwise be hidden from us within large quantities of data.

Analytics uses data and math to answer business questions, discover relationships, predicts unknown outcomes and automate decisions.

**Types of Analytics**

1. **DESCRIPTIVE:** This can be termed as the simplest form of analytics. The mighty size of big data is beyond human comprehension and the first stage hence involves crunching the data into understandable chunks.

The purpose of this analytics type is just to summarize the findings and understand what is going on.

1. **DIAGNOSTIC:** Diagnostic analytics is used to determine why something happened is the past. It is characterized by techniques such as drill-down, data discovery, data mining and correlations.

Diagnostic analytics takes a deeper look at data to understand the root causes of the events.

1. **REDICTIVE:** Predictive analytics is used to predict future outcomes. However, it is important to note that it cannot predict if an event will occur in the future; it merely forecasts what are the probabilities of the occurrence of the event.

A predictive model builds on the preliminary descriptive analytics stage to derive the possibility of the outcomes.

1. **PRESCRIPTIVE:** The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of “what-might- happen” analysis to help the user determine the best course of action to take. Prescriptive analysis is typically not just with one individual action, but is in fact a host of other actions.

***What is ANALYTICS?***

Data analytics is a broad term that defines the concept and practice (or, perhaps science and art) of all activities related to data.

**Types of Analysis**-

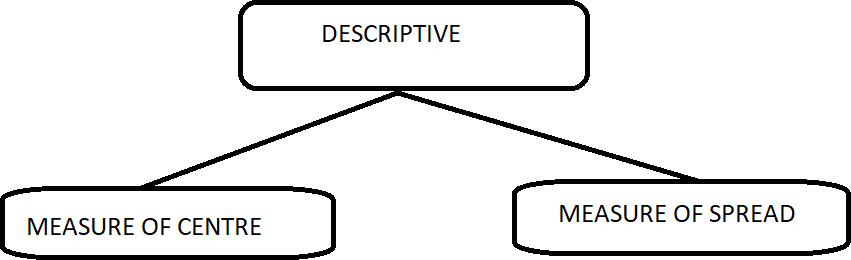
1. **Qualitative Analysis**

* It is mostly deals with generic data using text media
* It is also known as non-statistical analysis
* It deals with description
* Data can be observed but not measured
* Like – colors textures, smell, taste, beauty
* Qualitative = Quality

1. **Quantitative Analysis**

* It is the science of collecting and interpreting objects with numbers.
* It is also known as statistical analysis
* It is based on meaning derived from number.
* Data which can be measured.
* Like –length, height, area, weight, speed, time, temperature, cost age
* Quantitative= Quantity

## Descriptive Statistics



## Measure of Centre/ Measure of central tendency -

1. **Mean:** It is simply the average of all the data (salary) values. Add all the numbers then divide by the amount of numbers.

16000+15000+10000+12000+8000+18000=79000/6

The Mean is = 13166

**Examples:**

* What is mean of your last year expense?
* What is the average salary of your employee?
* What is the mean of your graduation score?
* What is mean of your bank accounts?

1. **Median:** It is the value in the middle when the data items are arranged in ascending order.

If you have two middle values then you should take average of both

**Examples:**

9,3,1,8,3,6

1,3,3,6,8,9

The Median is **4.5**

9,1,3,6,8

1,3,6,8,9

The Median is **6**

1. **Mode:** It is the most frequently occurring value in a series of data in case of no repeating values, there would be no mode.

**Example:**

9,3,1,8,3,6

The Mode is 3

* I want to know the Age mode of our class?
* If we want to know about Modi’s foreign visits?
* If mall owner wants to know his best seller product?
* Your restaurant visit will talk about your favorite restaurant/dish

## Measure of spread/ dispersion

As the name suggests, the measure of dispersion shows the scattering of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data.

Characteristics of Measure of Dispersion

* A measure of dispersion should be rigidly defined
* It must be easy to Calculate and understand
* Not affected much by the fluctuations of observations
* Based on all observation

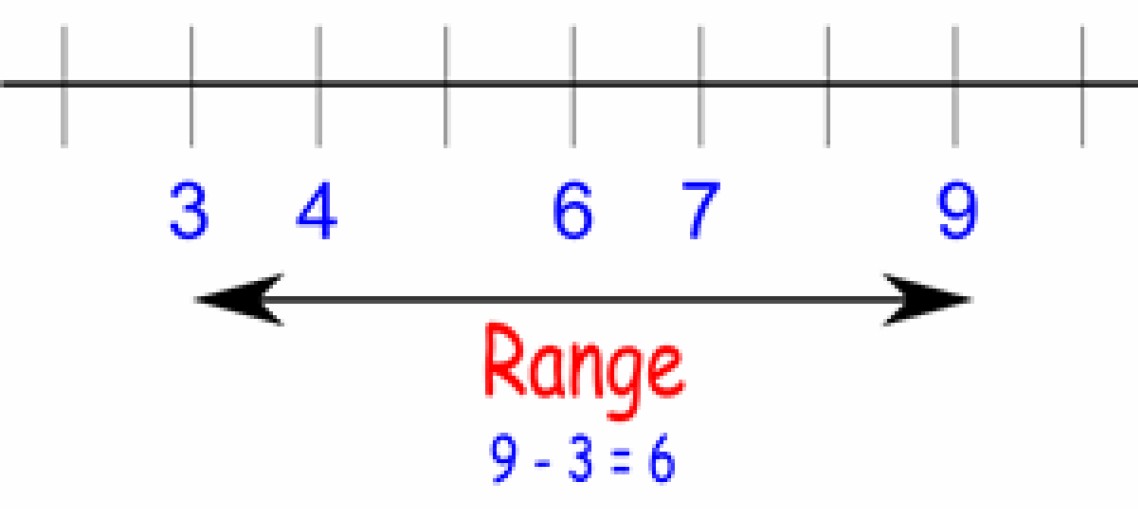
***Types of Measure of Dispersion***

1. **Absolute measure of dispersion:** The measures which express the scattering of observation in terms of distances i.e., range, quartile, deviation.
2. **Relative measure of dispersion:** The measures which express the variations in terms of the average of deviations of observation like mean deviation and standard deviation.

* **Range:** It is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set.

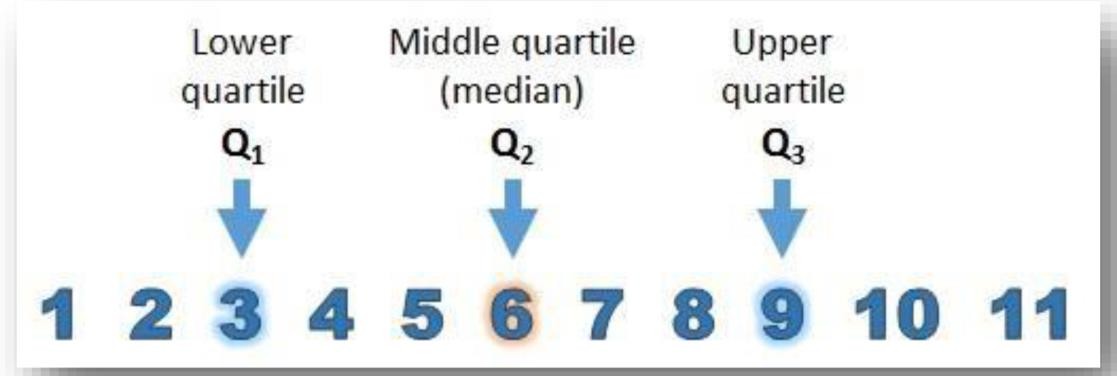
If X max and X min are the two extreme observations then.

Range = X max – X min



* **Quartile:** The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The second quartile, (Q2) is the median of the data set. The third quartile, (Q3) is the middle number between the median and the largest number.

A quartile divides a sorted data set into 4 equal parts, so that each part represents ¼ of data set.



* **Variance:** Variance is the average squared deviation from the mean of a set of data.

It is used to find the standard deviation.

* **Standard Deviation:** Standard Deviation shows the variation in data.

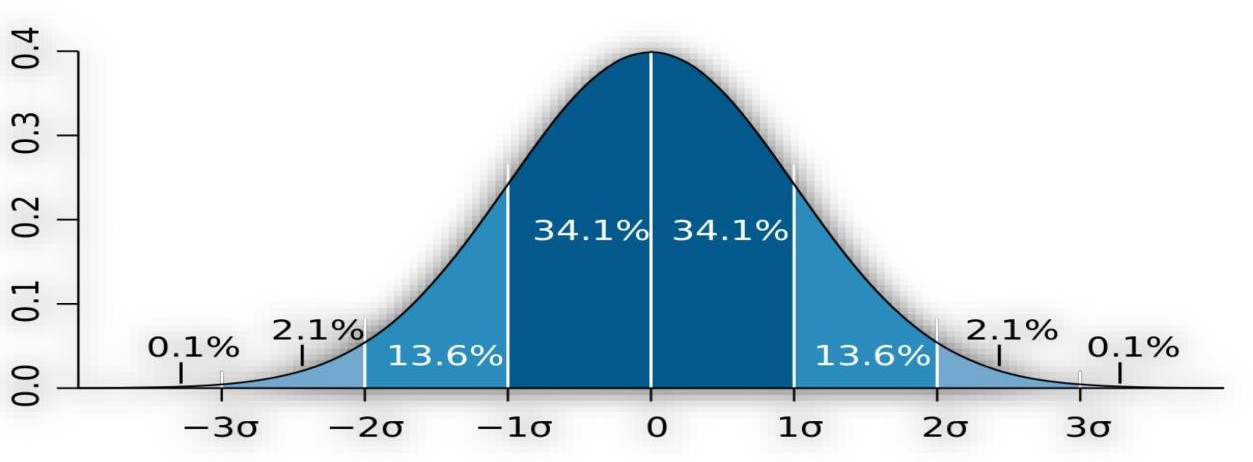
If the data is close together, the standard deviation will be small.

If the data is spread out, the standard deviation will be large.

Standard Deviation is often denoted by the lowercase Greek letter sigma.

The bell curve is commonly seen in statistics as a tool to understand standard deviation.

The following graph of a normal distribution represents a great deal of data in real life. The mean or average is represented by the Greek letter μ, in the center.



**How to find Standard Deviation**

For example, let’s find the standard deviation of the following data:

1,2,2,4,6

1. Calculate the mean of data: 15/5 = 3
2. Subtract the mean from each data value: -2, -1, -1, 1, 3
3. Square each of the new data value: 4,1,1,1,9
4. Sum these squared data values: 16
5. Divide this sum by (number of observations -1): 16 / (5-1) = 4
6. This number is Variance and Square root of this number is standard deviation: Sqrt (4) = 2

For instance, standard deviations of price data are frequently used as a measure of volatility; While monitoring some industrial process, if process indicators go beyond design standards then it may be troublesome hence variance/standard deviation can be used in such cases.

## What is Probability?

In the most literal sense, probability is the likelihood of the occurrence of an event.



* A card is drawn from a well shuffled pack of 52 cards. Find the probability of Ace
* Number of favorable outcomes i.e. ‘a jack’ is 4 out of 52 cards
* A king of red colour
* A card of diamond
* A black card
* Probability of an event= (Number of favorable outcomes)/ (Total Number of Possible Outcomes)
* P(A)=n(E)/n(S)
* Probability of Head and Tail

Mathematically, the probability that an event will occur is expressed as a number between 0 and 1.

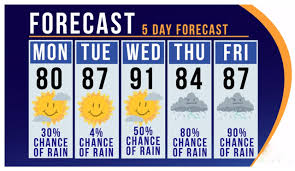
Notationally, the probability of event A is represented by P(A).

* If P(A) equals zero, event A will almost definitely not occur.
* If P(A) is close to zero, there is only a small chance that event A will occur.
* If P(A) equals 0.5, there is a 50-50 chance that event A will occur.
* If P(A) is close to one, there is a strong chance that event A will occur.
* If P(A) equals one, event A will almost definitely occur.

### Real Time Example of Probability in Data Science -

* Weather Forecasting- Before planning for an outing or a picnic, we always check the weather forecast. Suppose it says that there is a 60% chance that rain may occur.
* Batting average represents how many runs a batsman would score before getting out. For example, if a batsman had scored 40 runs out of 100 from boundaries in the previous match. Then, there is a chance that he would score 40% of his runs in the next match from boundaries.
* Insurance-For example, you are an active smoker, and chances of getting lungs disease are higher in you. So, instead of choosing an insurance scheme for your vehicle or house, you may go for your health insurance first, because the chance of your getting sick are higher.
* Lottery Tickets-In a typical Lottery game, each player chooses six distinct numbers from a particular range. If all the six numbers on a ticket match with that of the winning lottery ticket, the ticket holder is a Jackpot winner- regardless of the order of the numbers. The probability of this happening is 1 out of 10 lakh.





## Conditional Probability-

Conditional probability is the probability of one event occurring with some relationship to one or more other events.

**Example:**

Suppose you are asked to draw a card out of pack of 52 cards.

You drew a red card.

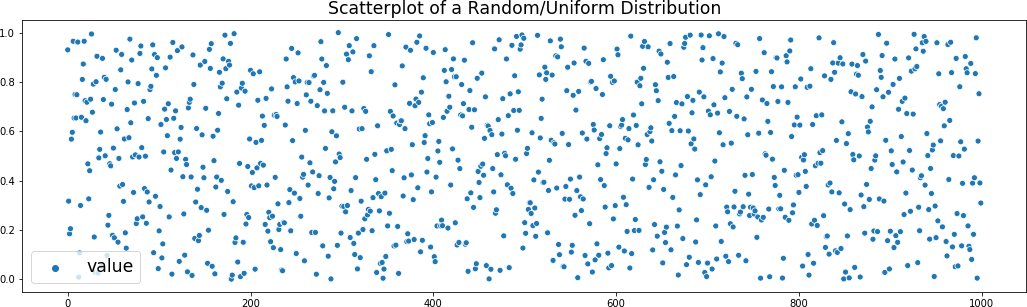
What is the probability that it’s a four?

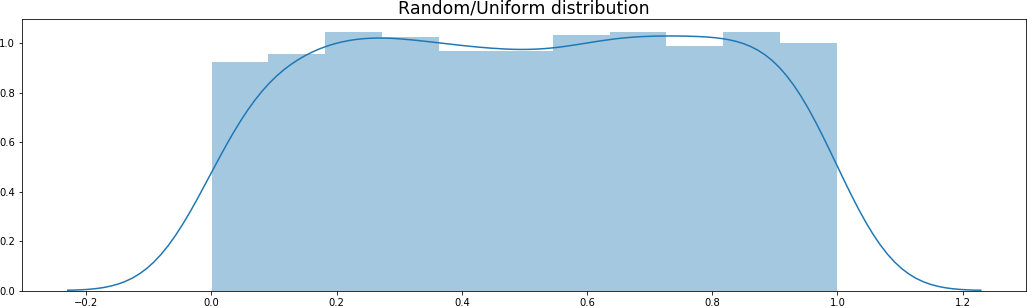
Probability of getting red card given it’s a 4=p(four|red)=4/52=2/26=1/13.

So out of the 52 red cards (given a red card), there are 4 fours. 

## Type of Probability Distribution-

1. **Uniform distribution:** It is fairly simple. Every value has a change of incidence that is equal. The distribution is thus made up of random values with no trends in them.





A uniform distribution can be used for every case in which any result in a sample space is equally possible. One instance of this is rolling a single standard die in a discrete situation. A total of six sides of the die are open, and each side is equally likely to be rolled face up. For this distribution, the probability histogram is rectangular, with six bars that each have a height of 1/6.

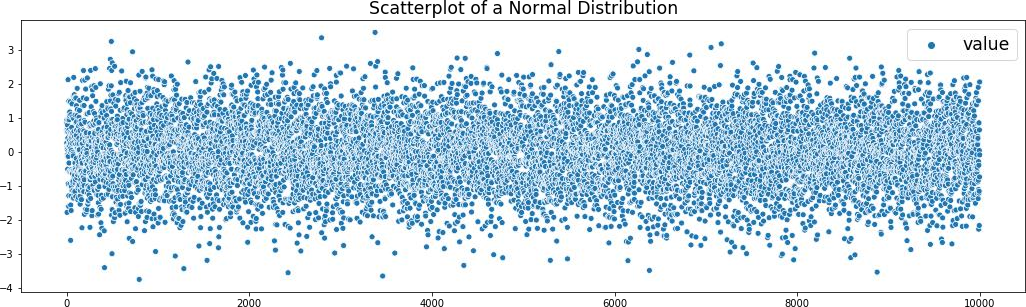
1. **Normal Distribution**: The "Bell Curve" is a Normal Distribution and some data that follows it closely, but not perfectly (which is usual). It is often called a "Bell Curve“because it looks like a bell.

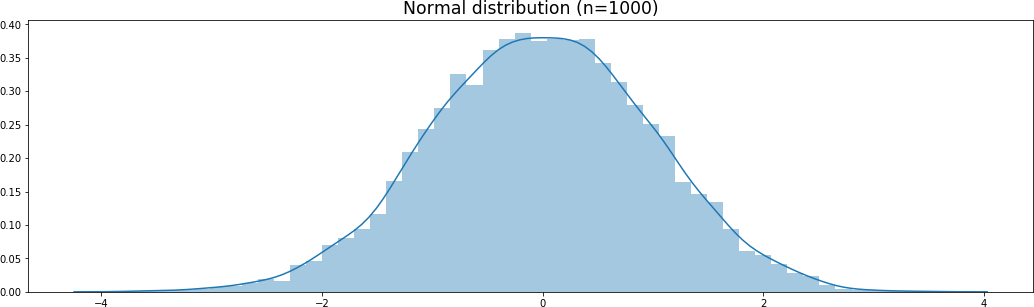
The Normal Distribution has:

* Mean = Median = Mode
* Symmetry about the center
* 50% of values less than the mean and 50% greater than the mean

**Example-**

Most of us have heard about the rise and fall in the prices of the shares in the stock market. our parents or in the news about falling and hiking in the price of the shares. These changes in the log values of Forex rates, price indices, and stock prices return often form a bell-shaped curve. For stock returns, the standard deviation is often called volatility. If returns are normally distributed, more than 99 percent of the returns are expected to fall within the deviations of the mean value. Such characteristics of the bell-shaped normal distribution allow analysts and investors to make statistical inferences about the expected return and risk of stocks.





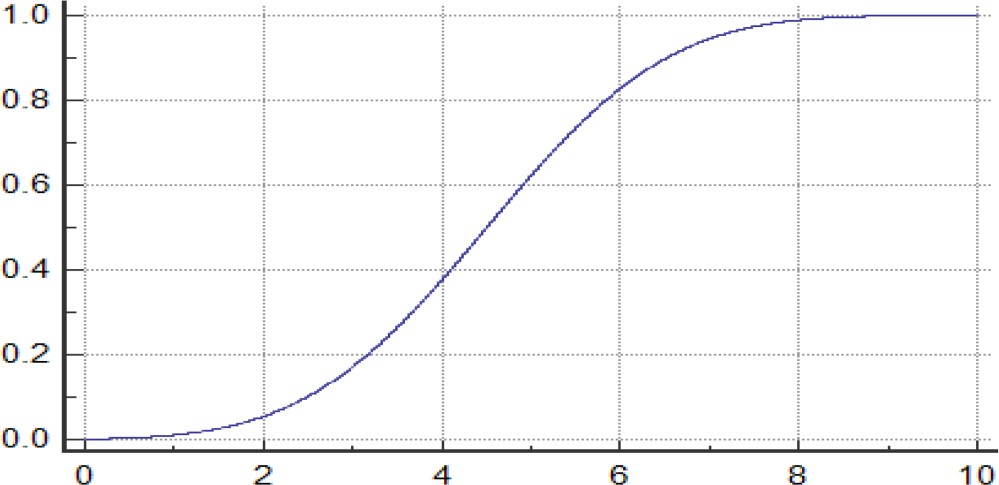
Binomial Distribution-is a type of distribution that has two possible outcomes

(the prefix “bi” means two, or twice).

For example, a coin toss has only two possible outcomes:

heads or tails

taking a test could have two possible outcomes: pass or fail.



1. **Binomial Distribution**: It is considered a distribution where only two results are possible, such as success or defeat, gain or loss, win or lose, and where the chance of success and failure is the same for all the experiments.

The outcomes may not necessarily be equally likely. So, if the likelihood of success in an experiment is 0.2, then it is straightforward to measure the probability of failure as q = 1- 0.2 = 0.8.

A Binomial experiment is an experiment that has only two possible outcomes when repeated n number of times.

N and p are the parameters of a binomial distribution, where n is the total number of trials and p is the likelihood of each trial's success.

**Following are the properties of Binomial Distribution:**

* Every trial is independent
* There are only two possible outcomes in a trial- either a success or a failure.
* A total number of n identical trials are conducted.
* For every experiment the probability of success and failure is the same.

**Example-**

Assume that a casino created a new game in which participants are able to place bets on the number of heads or tails in a specified number of coin flips. Assume a participant wants to place a $10 bet that there will be exactly six heads in 20 coin flips. The participant wants to calculate the probability of this occurring, and therefore, they use the calculation for the binomial distribution.

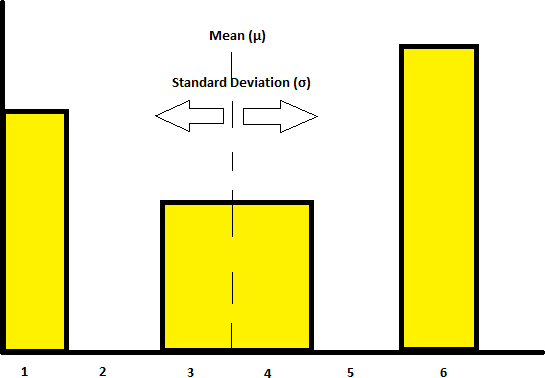
## Central Limit Theorem

In stats, CLT has a central principle that allows you to use data to analyze your hypotheses, even with lacking information, so it is one of the foundations of hypothesis testing, an important statistical decision- making.

Central Limit Theorem states that when the large sample size has a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.

Let's take a random 6-sided dice distribution, the probability distribution

function of which is with mean μ and standard deviation σ.



From the image given above, we can see that this dice cannot get 2 and 5.

Let’s take samples from this distribution of sample size 4, that is we’ll

take 4 random samples from the population.

Sample1(S1) = (1,1,3,6)

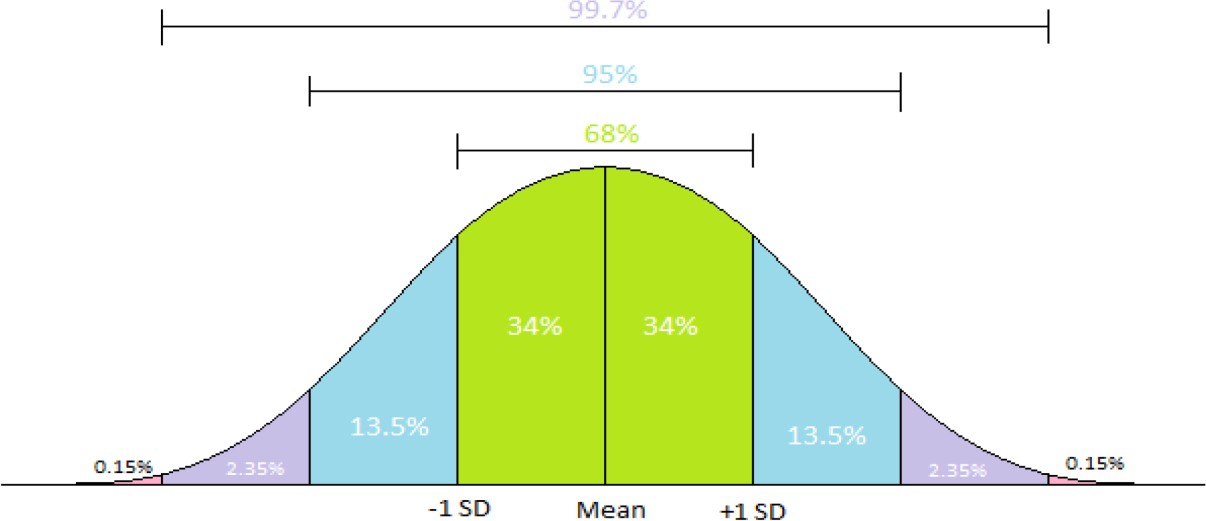
its mean is x1= (1+1+3+6)/4 = 2.75 S2 = (3,4,3,1)

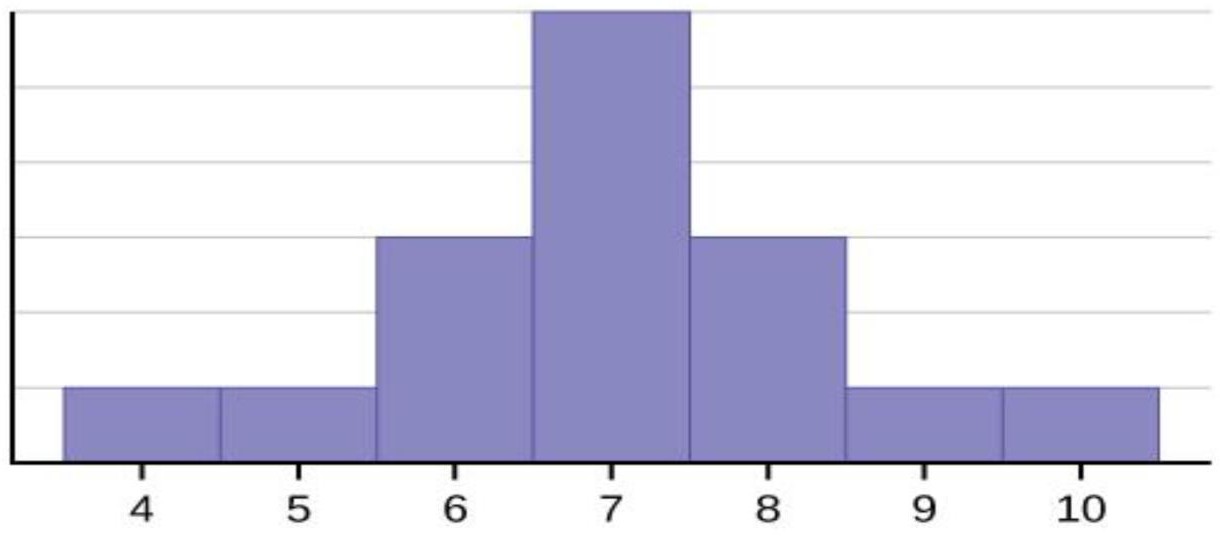
x2 = (3+4+3+1)/4 = 2.75 S3 = (1,1,6,6)

x3 = (1+1+6+6)/4 = 3.5

***Skewness-***

* It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.
* It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0



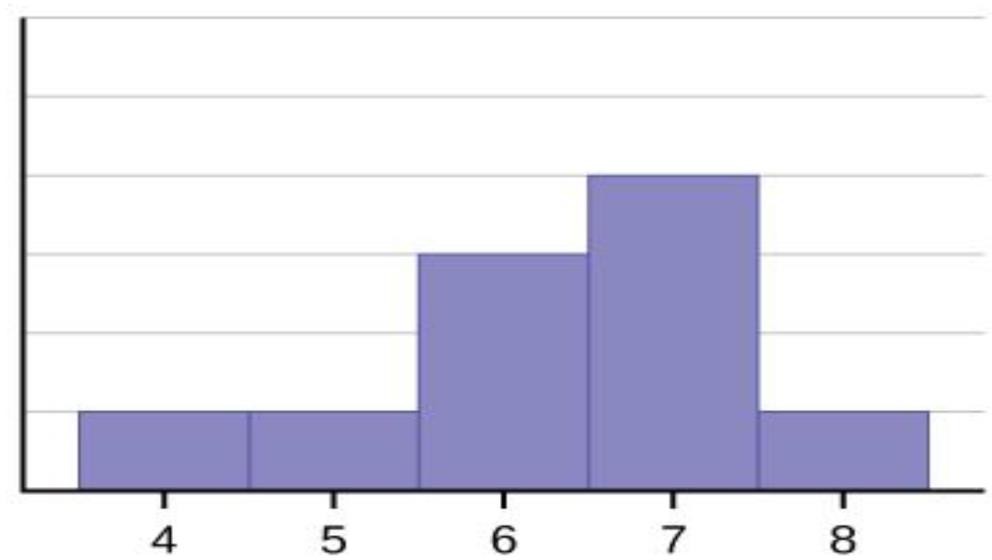
**No Skewness-**

Consider the following data set.

* 4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10
* Mean=Mode=Median
* Mean =7
* Median=7
* Mode=7

### Left/ Negative skewness-

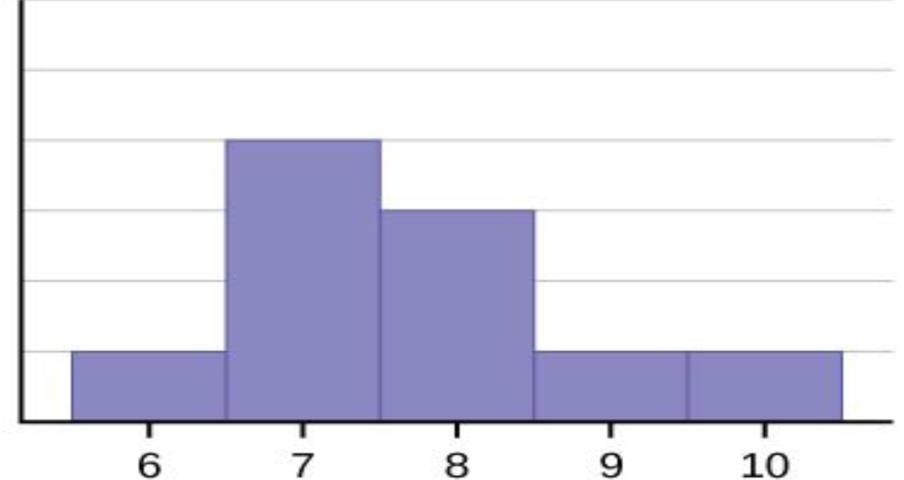
Consider the following data set.

* 4; 5; 6; 6; 6; 7; 7; 7; 7; 8
* Mode>Median>Mean
* Mean =6.3
* Median=6.5
* Mode=7

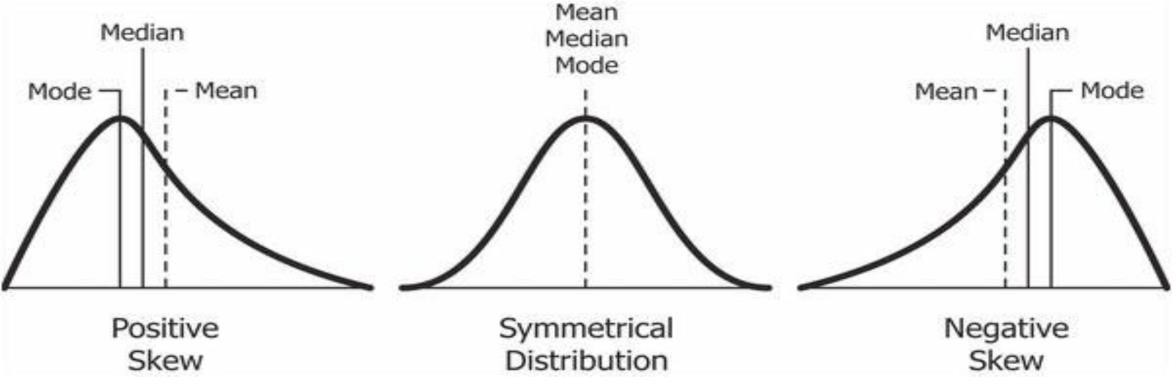
### Right/Negative Skewness-

Consider the following data set.

* 6;7;7; 7; 7; 8; 8;8;9;10
* Mean =7.7
* Median=7.5
* Mode=7

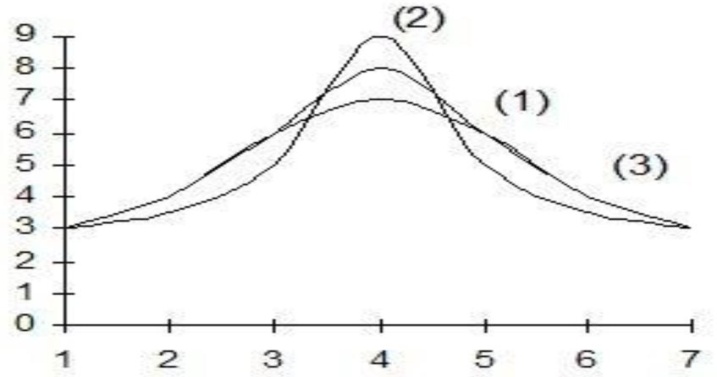


* Left or Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side.
* Right or Positive Skewness means when the tail on the right side of the distribution is longer or fatter.



## Kurtosis-

A measure of the peakness or convexity of a curve is known as Kurtosis.



* Curve (1) is known as mesokurtic (normal curve);
* Curve (2) is known as leptocurtic (leading curve)
* Curve (3) is known as platykurtic (flat curve).

## Outliers-

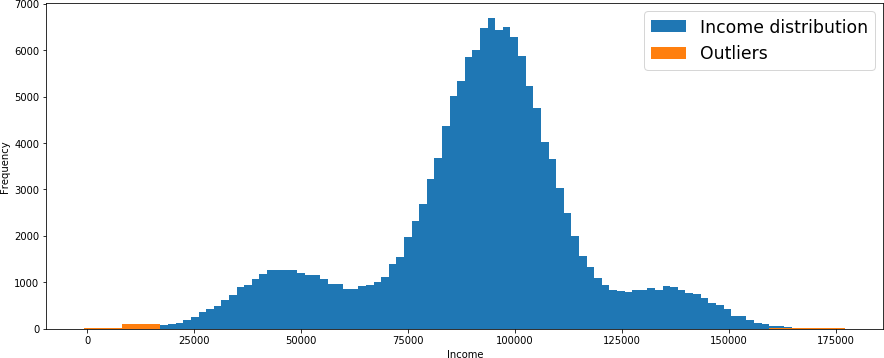
Outliers have many applications in determining fraud and potential trends in the market. The first definition was given by Grubbs in 1969 as “An outlying observation, or outlier is one that appears to be very different from other members of the sample in which it occurs”.

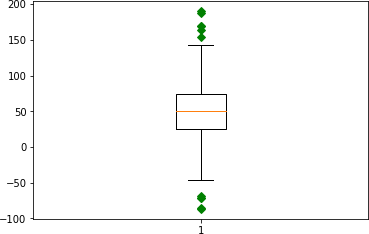
The concept of outliers is closely associated with robustness. Outliers are “abnormal” observations in the sample that seem very unlikely for the assumed distribution model or are remarkably different from the rest of sample observations. Outliers can be originated by measurement errors, exceptional circumstances, changes in the data generating process, etc.

Outliers are quite different from the noise data.

* Noise is random error or variance in a measured variable
* Noise should be removed before outlier detection

An outlier may indicate an experimental error, or it may be due to variability in the measurement.





## Covariance and Correlation

Covariance and correlation are two opposed terms and are used in regression analysis. Covariance shows how the two variables differ, whereas correlation shows how the two variables are related.

## Covariance

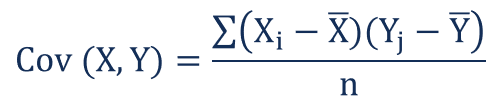
Covariance refers to relationship between two random variables in which a change in the one reflects a change in other variable.

The range of covariance is from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

The greater this number, the more reliant the relationship.

Covariance is great for defining the type of relationship, but it's terrible for interpreting the magnitude. And, here comes the concept of correlation.

**Formula for covariance:**



Where:

* Xi – the values of the X-variable
* Yj – the values of the Y-variable
* X̄ – the mean (average) of the X-variable
* Ȳ – the mean (average) of the Y-variable
* n – the number of data points

## Correlation

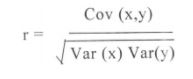
Correlation refers to how two variables move in relation to one another.

The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1. When it is close to 0 this means that there is little relationship between the variables and when it is farther away from 0 in positive or negative direction, greater is the relationship between the two variables.

Correlation is a measure that determines the degree to which two or more random variables move in sequence. When movement of one variable reciprocates the movement of another variable in some way or another, the variables are said to be correlated.

When variables move in the same direction, they are said to be positively correlated and when they move in the opposite direction, they are said to be negatively correlated.

**Formula for correlation:**



where,

var(x) = variance of x

var(y) = variance of Y

## Types of Correlation

1. **Positive correlation**: A positive correlation would be 1. This means the two variables moved either up or down in the same direction together.
2. **Negative correlation**: A negative correlation is -1. This means the two variables moved in opposite directions.
3. **Zero or no correlation**: A correlation of zero means there is no relationship between the two variables. In other words, as one variable moves one way, the other moved in another unrelated direction.

## Difference between Correlation& Covariance

|  |  |  |
| --- | --- | --- |
|  | **Correlation** | **Covariance** |
| **Meaning** | Correlation is an indicator of how strongly these 2 variables are related, provided other conditions are constant. The maximum value is +1, denoting a perfect dependent relationship. | Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. |
| **Relationship** | Correlation provides a measure of covariance on a standard scale. It is deduced by dividing the calculated covariance with standard deviation. | Correlation can be deduced from a covariance. |
| **Values** | Correlation is limited to values between the range  -1 and +1. | The value of covariance lies in the range of -∞ and  +∞. |
| **Scalability** | Correlation is not affected by a change in scales or multiplication by a  constant. | Affects covariance |
| **Units** | Correlation is a unitless absolute number between -1 and +1,  including decimal values. | Covariance has a definite unit as it is deduced by the multiplication of two  numbers and their units |

## Similarities between Correlation& Covariance

Correlation and Covariance both measures the linear relationships between two variables. It means that when the correlation coefficient is zero, the covariance is also zero. But, when it comes to making a choice between covariance and correlation to measure relationship between variables, correlation is preferred over covariance because it does not get affected by the change in scale.