FINGERTIPS

Module 4

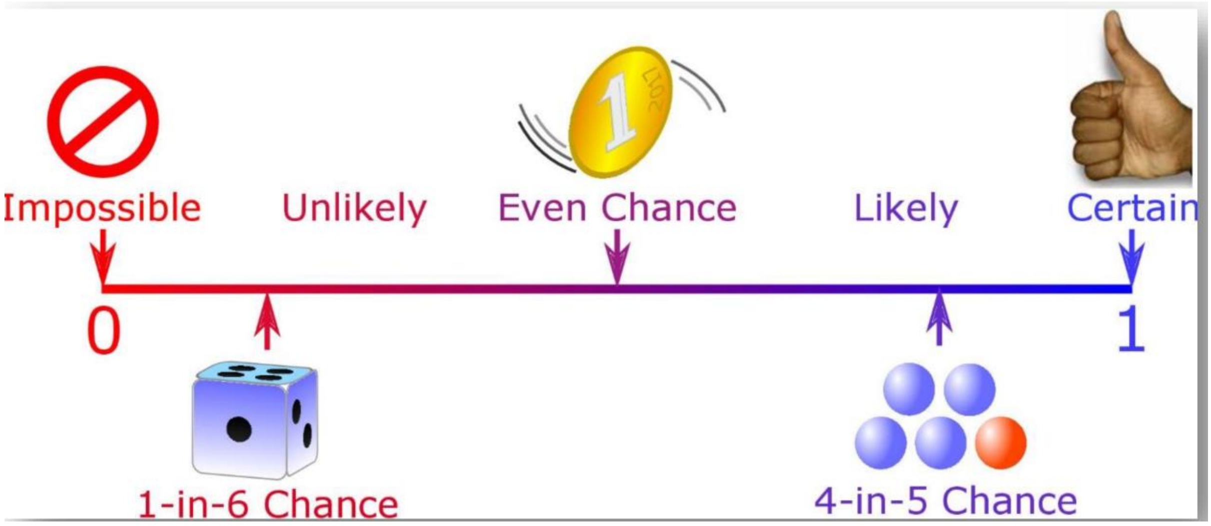
Probability & Probability Distribution

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.

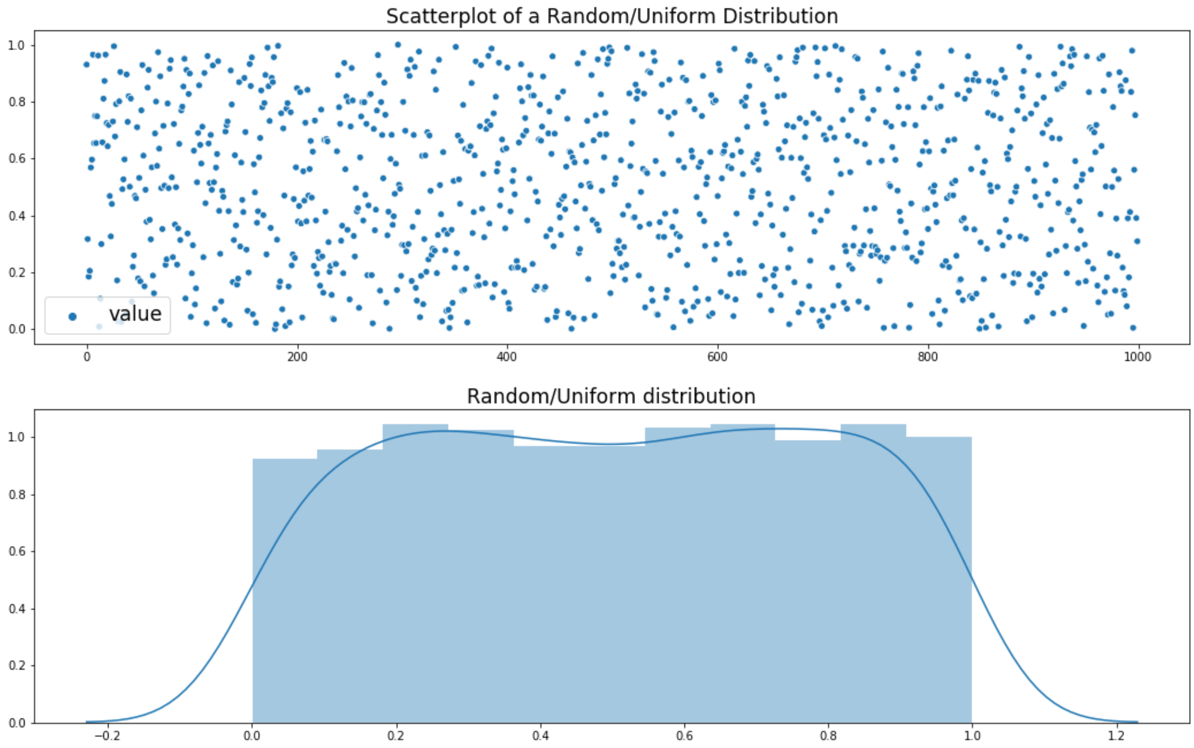
- Batting average represents how many runs a batsman would score before getting out. For example, if a batsman had scored 40 runs out of 100 from boundaries in the previous match. Then, there is a chance that he would score 40% of his runs in the next match from boundaries.

- **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**.**

Type of Probability Distribution-

Uniform distribution 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.

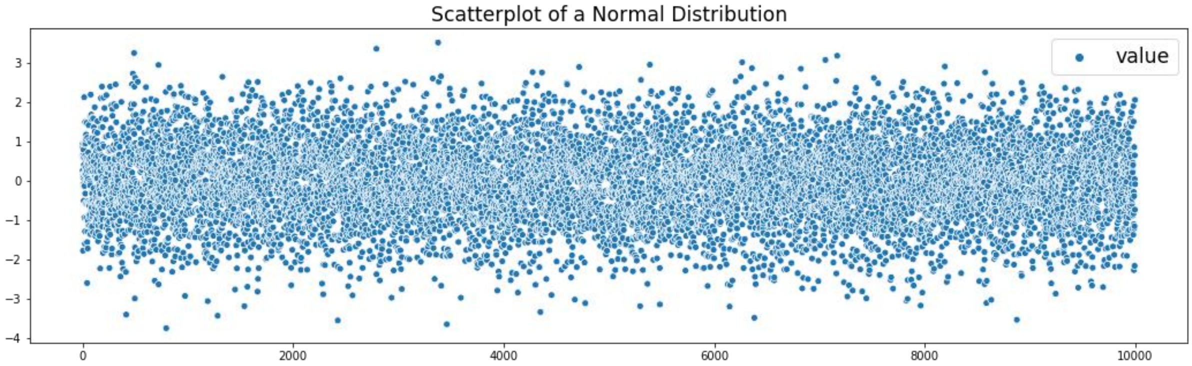
Normal Distribution-The "Bell Curve" is a Normal Distribution and some data that follows it closely, but not perfectly (which is usual). It is often called a "Bell Curve“because it looks like a bell.

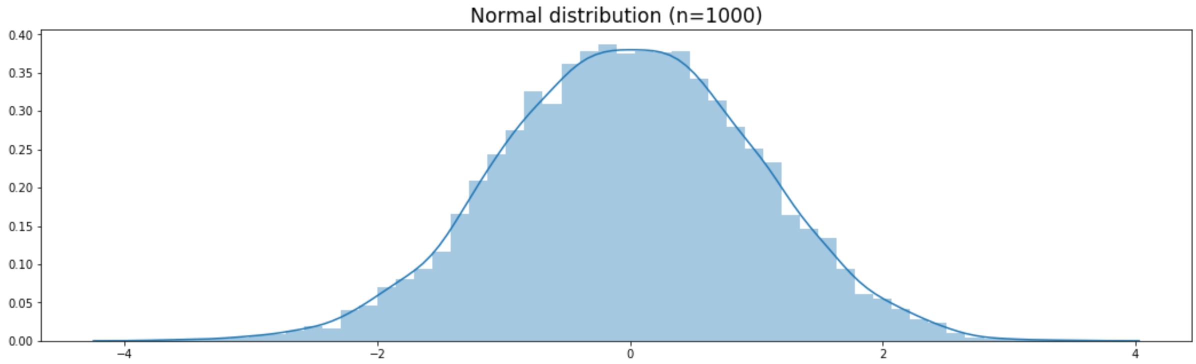
The Normal Distribution has:

- 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.



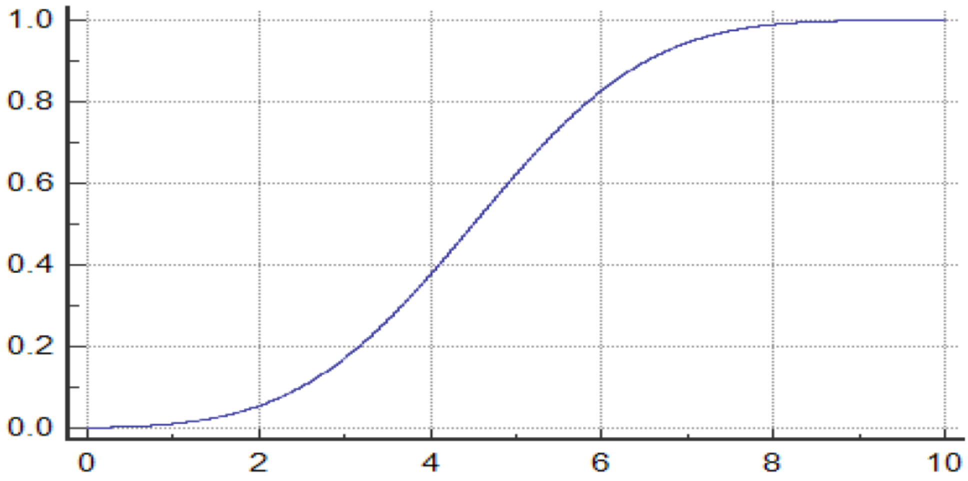


Binomial Distribution-is a type of distribution that has two possible outcomes (the prefix "bi" means two, or twice).

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

- heads or tails

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



- A Binomial Distribution 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

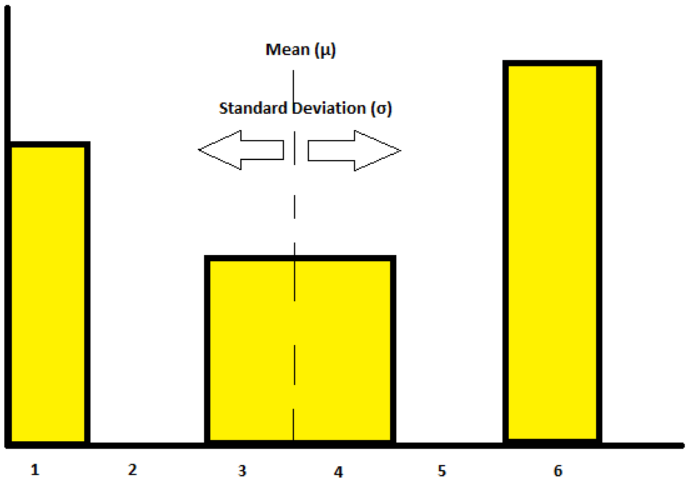
1. Every trial is independent
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. For every experiment the probability of success and failure is the same.

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.
- 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

Conditional Probability-

- Conditional probability is the probability of one event occurring with some relationship to one or more other events
- Event A-That it is raining outside 30% Chance of Raining today.
- Event B- you will need to go outside, and that has a probability of 0.5 (50%).

Bayes Theorem

Bayes theorem is a way to figure out conditional probability. Conditional probability is the probability of an event happening, given that it has some relationship to one or more other events. ... In a nutshell, it gives you the actual probability of an event given information about tests.

Given a set of events A1,A2, … ,An, the probability of all of them occurring simultaneously is called probability chain rule, and it is

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)\ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}) = \prod_{k=1}^{n} P(A_k|\bigcap_{j=1}^{k-1} A_j)$$

Or alternatively, in two events form,

$$P(A \cap B) = P(A)P(B|A)$$

Total Probability Rule

Given A1, … ,An mutually exclusive (disjoint) events whose union is the whole of the sample space (partition) and assume P(Ai) > 0 for every i. For every event B we have

$$P(B) = P(A_1 \cap B) + \cdots + P(A_n \cap B) = \prod_{k=1}^{n} P(A_k|\bigcap_{j=1}^{k-1} A_j)$$

Bayes Theorem

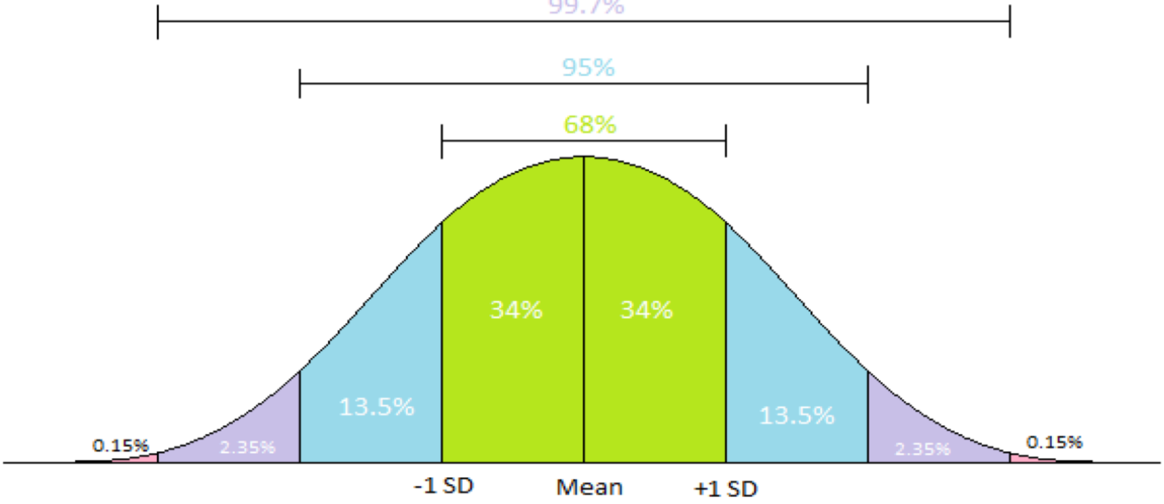
Knowing above, we can write

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(A_1 \cap B) + \cdots + P(A_n \cap B)}$$

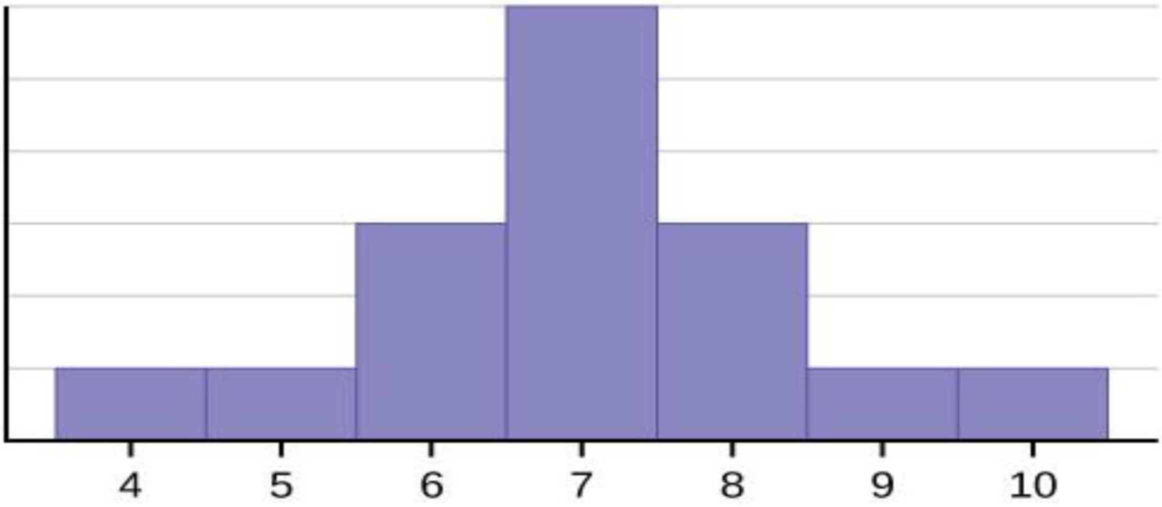
Skewness-

- 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-

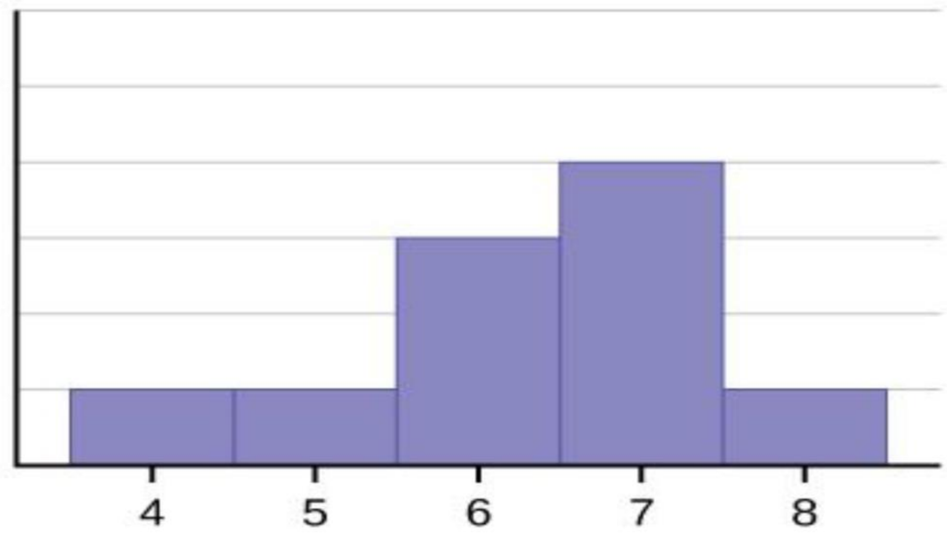


Consider the following data set.

- 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-

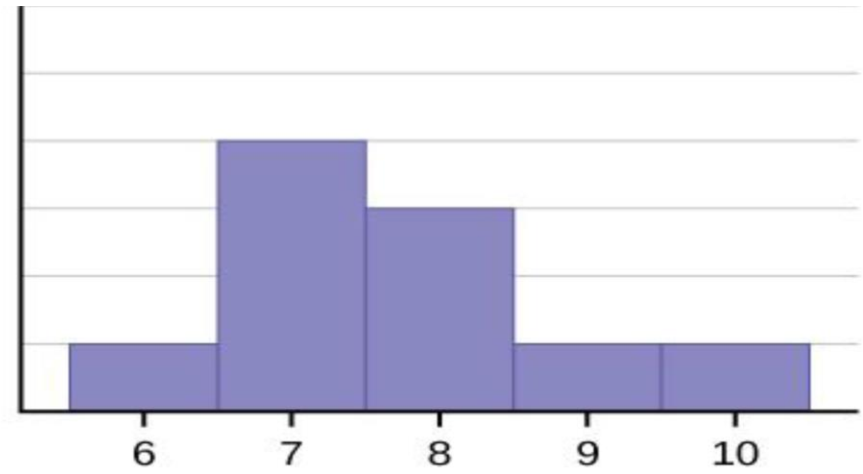
- Consider the following data set.
- 4; 5; 6; 6; 6; 7; 7; 7; 7; 8



- Mode>Median>Mean
- Mean =6.3
- Median=6.5
- Mode=7

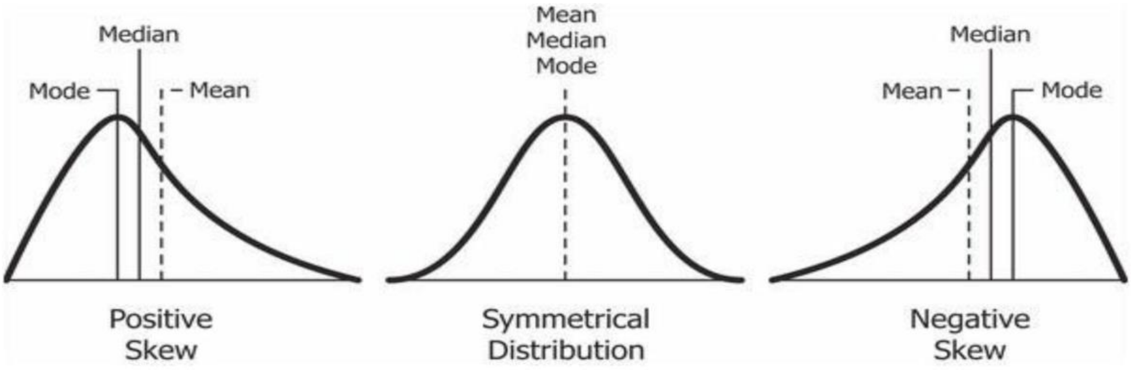
Right/Negative Skewness-

- Consider the following data set.
- 6;7;7; 7; 7; 8; 8;8;9;10



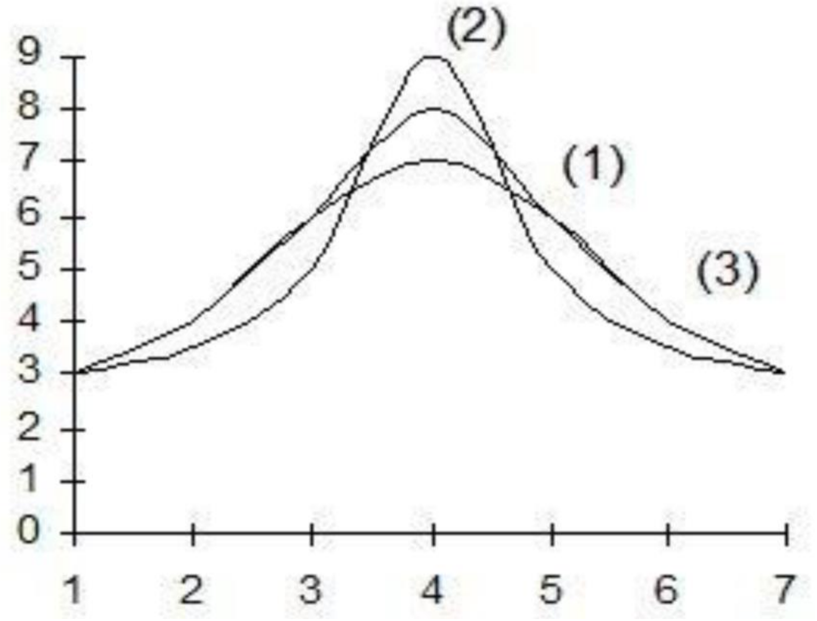
- Mean =7.7
- Median=7.5
- Mode=7

- Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side.
- 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).

Degrees of Freedom in Statistics-

Unfortunately, you have constraints. You have only 7 hats. Yet you want to wear a different hat every day of the week.

DF=7-1 = 6 days of “hat” freedom-in which the hat you wore could vary!

Degrees of freedom are often broadly defined as the number of "observations" (pieces of information) in the data that are free to vary when estimating statistical parameters.

Time Series-

One definition of a time series is that of a collection of quantitative observations that are evenly spaced in time and measured successively**.**

Example-

Monthly rainfall

Monitoring Heart rate

Hourly Temperature

Daily Closing price of a company Stock

Goals of time series analysis:

1. Descriptive: Identify patterns in correlated data—trends and seasonal variation
2. Explanation: understanding and modeling the data
3. Forecasting: prediction of short-term trends from previous patterns
4. Intervention analysis: how does a single event change the time series?
5. Quality control: deviations of a specified size indicate a problem