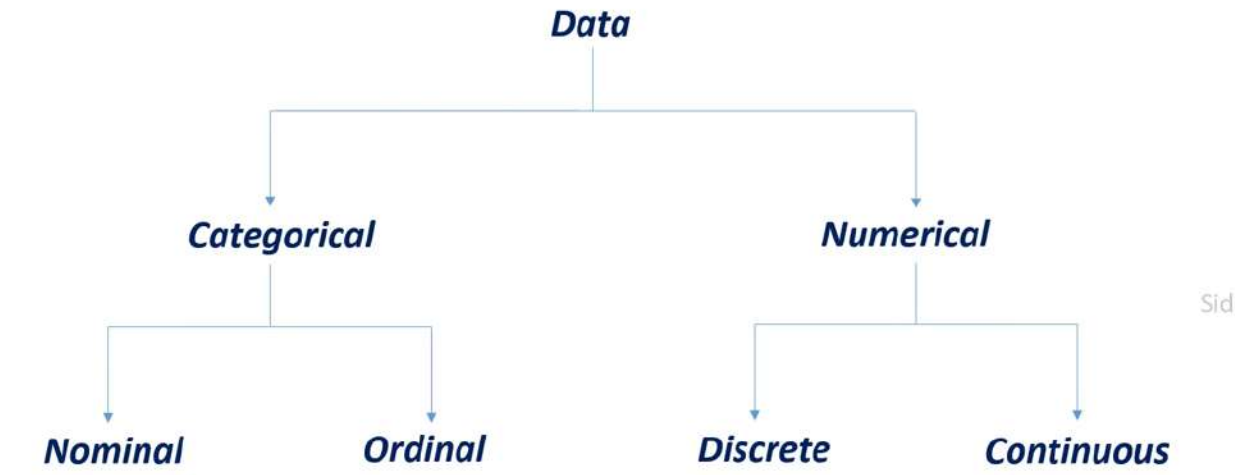
Types of Data



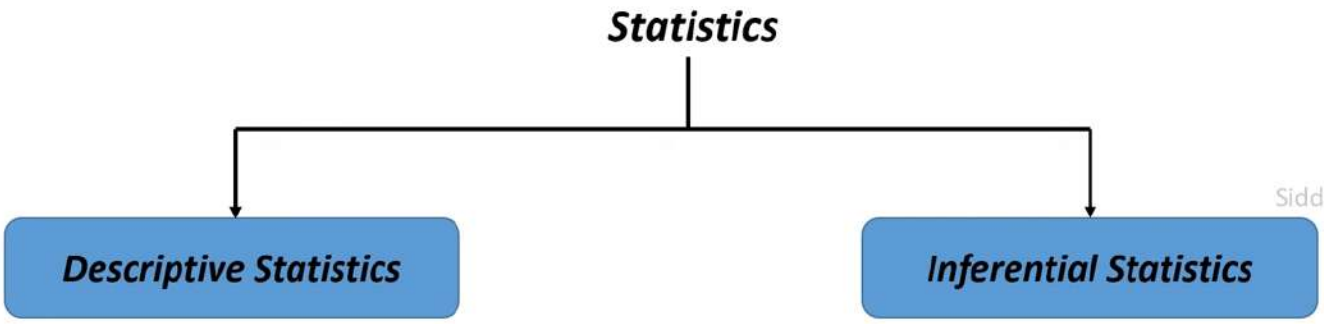
Nominal data *is a classification of categorical variables, that do not provide any quantitative value.*

Ordinal data *are the type of data in which the values follow a natural order.*

Discrete Data *are the type of data that can only take certain values.*

Continuous data *can have almost any numeric value. Unlike Discrete data, they can have decimal values.*

Types of Statistics



- Describe the data
- Data Analysis

- Get inferences about the data
- Data Science

Types of Statistical Studies

Statistical Study

Sample Study

*A **sample study** is a study which is carried out on a sample which represents the total population.*

Observational Study

*An **observational study** is a study where we simply collect and analyze data. We won't inject any changes. We just observe the correlation in the data.*

Experimental Study

*An **experimental study** is a **study** in which conditions are controlled and manipulated by the experimenter.*

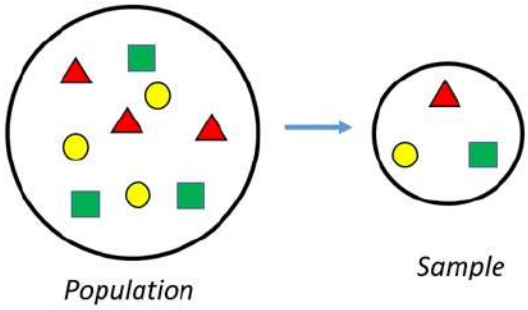
Types of Statistics

1. Descriptive Statistics:

Descriptive statistics *are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures.*

2. Inferential Statistics:

Inferential statistics *takes data from a sample and makes inferences and predictions about the larger population from which the sample was drawn.*



Descriptive Statistics

2 important measures of Descriptive Statistics:

1. Measure of Central Tendencies (Mean, Median, Mode)
2. Measure of Variability (Range, Standard Deviation, Variance)

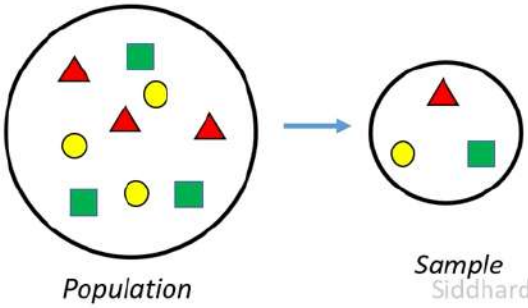
Siddhardhan

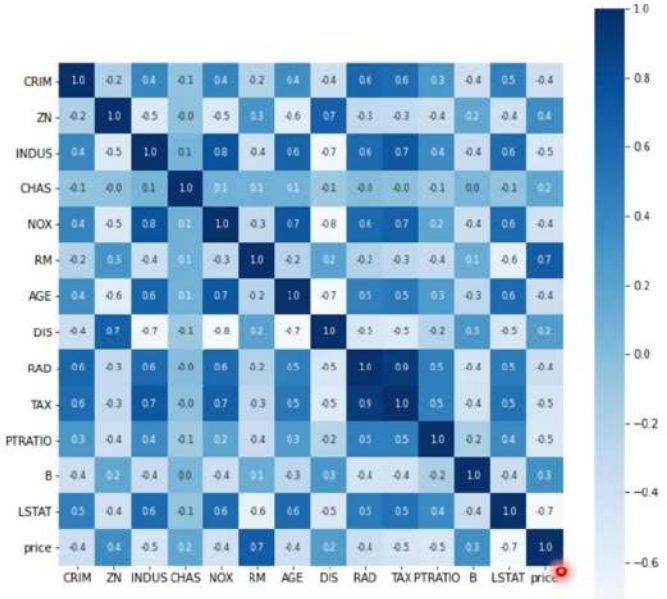
Descriptive Statistics of House Price Dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	price
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

Inferential Statistics

Inferential statistics *takes data from a sample and makes inferences and predictions about the larger population from which the sample was drawn.*

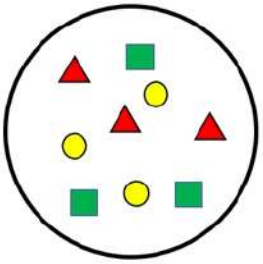




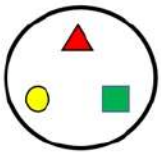
Correlation of House Price Data

1. Sample Study

*A **sample study** is a study which is carried out on a sample which represents the total population.*



Population

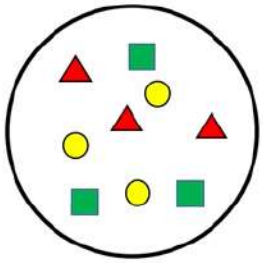


Sample

Average Blood Sugar Level = ?

2. Observational Study

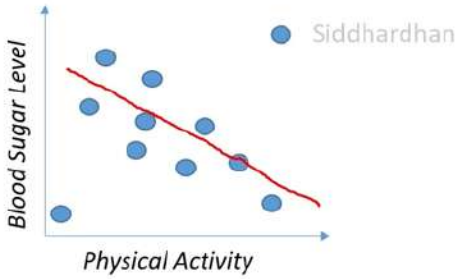
*An **observational study** is a study where we simply collect and analyze data. We won't inject any changes. We just observe the correlation in the data.*



Population

Relation between:

1. Blood Sugar Level
2. Physical Activity

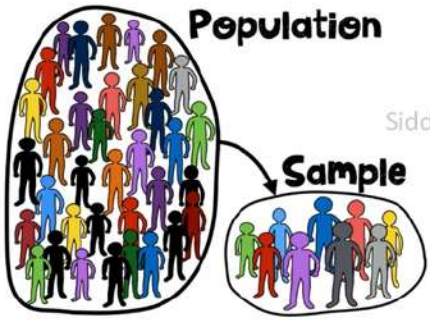


Inference: Blood Sugar Level & Physical Activity are Negatively Correlated

Siddhardhan

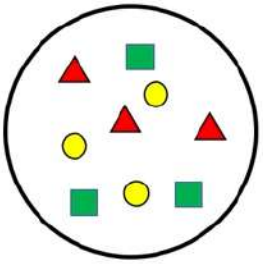
Population & Sample - Sampling Techniques

Math for Machine Learning

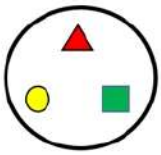


1. Sample Study

*A **sample study** is a study which is carried out on a sample which represents the total population.*



Population



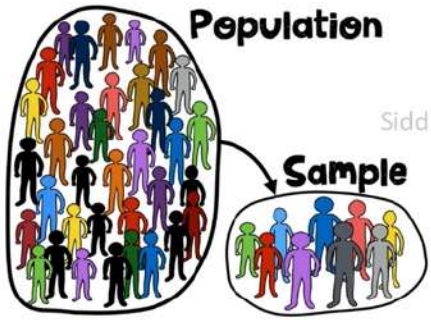
Sample

Average Blood Sugar Level = ?

Types of Sampling Techniques

Sampling Techniques:

- Simple Random Sampling
- Systematic Sampling
- Stratified Random Sampling
- Cluster Sampling



(Probability Sampling Techniques)

(Non-Probability Sampling Techniques)

Simple Random Sampling

In **Simple Random Sampling**, the sample is randomly picked from a larger population. Hence, all the individual datapoints has an equal probability to be selected as sample data.

Example: Employee survey in a company

Pros:

1. No sample Bias
2. Balanced Sample
3. Simple Method of sampling
4. Requires less domain knowledge

Cons:

1. Population size should be high
2. Cannot represent the population well sometimes

Systematic Sampling

In **Systematic Sampling**, the sample is picked from the population at regular intervals. This type of sampling is carried out if the population is homogeneous and the data points are uniformly distributed

Example: Selecting every 10^{th} member from a population of 10,000

Pros:

1. Quick & easy
2. Less bias
3. Even distribution of data

Cons:

1. Data manipulation risk
2. Requires randomness in data
3. Population should not have patterns.

Cluster Sampling

Cluster Sampling is carried out on population that has inherent groups. This population is subdivided into **clusters** and then random clusters are taken as sample.

Example: Smartphone sales in randomly selected states

Pros:

1. Requires only fewer resources
2. Reduced Variability
3. Advantages of both Random sampling and Stratified Sampling

Cons:

1. Cannot be performed on populations without natural groups
2. Overlapping data points
3. Can't provide a general insight for the entire population

Siddhardhan

Measure of Central Tendencies: Mean, Median & Mode

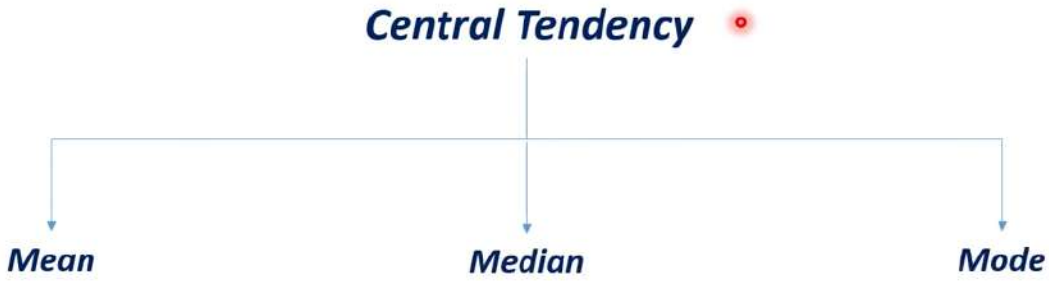
Math for Machine Learning

Central Tendency

Central Tendency:

*A measure of **central tendency** is a value that represents the center point or typical value of a dataset. It is a value that summarizes the data.*

Siddhardhan



Central Tendencies

Mean

Mean *or arithmetic mean is the sum of values divided by the number of values.*

$$M = \frac{\Sigma x}{N}$$

Heights

160
172
165
168
174

$$\frac{160+172+165+168+174}{5}$$

Mean = 167.8

Median

*The **median** is the **middle** value in the list of numbers. To find the median, the numbers have to be listed in numerical order from smallest to largest.*

160 165 168 172 174

160 165 168 172 174 176

$$\frac{168+172}{2} = 170$$

Median = 170

Mode

*The **mode** is the value that occurs most often. If no number in the list is repeated, then there is no mode for the list.*

Siddhardhan

Heights

160
172
160
168
174

Mode = 160

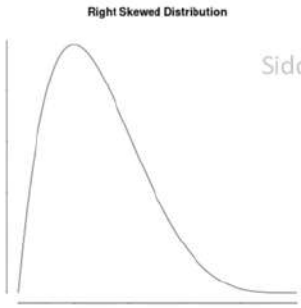
Central Tendencies in Data Pre-Processing

Central Tendencies are very useful in **handling the missing values** in a dataset

Mean : Missing values in a dataset can be replaced with **mean** value, if the data is uniformly distributed.

Median : Missing values in a dataset can be replaced with **median** value, if the data is skewed.

Mode : Missing values in a dataset can be replaced with **mode** value, if the data is skewed. Missing categorical values can also be replaced with **mode** value.

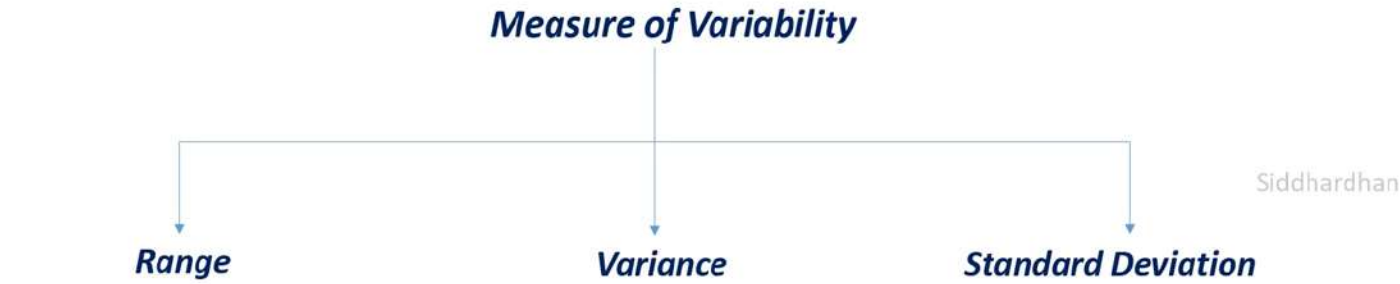


Siddhardhan

Measure of Variability: Range, Variance & Standard Deviation

Math for Machine Learning

Measure of Variability



*The **range** of a set of data is the difference between the largest and smallest values. It can give a rough idea about the distribution of our dataset.*

***Variance** is a measure of how far each number in the set is from the mean and therefore from every other number in the dataset.*

Standard Deviation is the square root of Variance. Standard deviation looks at how spread out a group of numbers is from the mean.

Measure of Variability

Measure of Variability

Range

*The **range** of a set of data is the difference between the largest and smallest values. It can give a rough idea about the distribution of our dataset.*

$$Range = Max\ value - Min\ Value$$

Variance

***Variance** is a measure of how far each number in the set is from the mean and therefore from every other number in the dataset.*

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

Standard Deviation

Standard Deviation is the square root of Variance. Standard deviation looks at how spread out a group of numbers is from the mean.

$$SD = \sqrt[2]{\sigma}$$

Range ; Variance ; Standard Deviation

-5, 0, 5, 10, 15,

$$Mean = \frac{-5+0+5+10+15}{5} = 5$$

Range = 15 – (-5) = 20

$$Variance = \frac{(-5-5)^2+(0-5)^2+(5-5)^2+(10-5)^2+(15-5)^2}{5}$$

$$Variance = 50$$

$$Standard\ Deviation = 7.1$$

3, 4, 5, 6, 7

$$Mean = \frac{3+4+5+6+7}{5} = 5$$

Range = 7 – 3 = 4

$$Variance = \frac{(3-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-5)^2}{5}$$

$$Variance = 2$$

$$Standard\ Deviation = 1.4$$

Siddhardhan

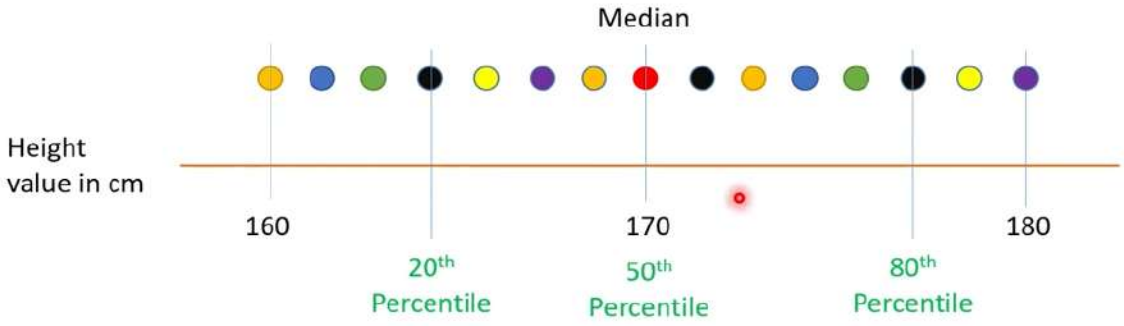
Percentiles & Quantiles

Math for Machine Learning

Percentiles

Percentile is a value on a scale of 100 that indicates the percent of a distribution that is equal to or below it.

Dataset with Height of 15 people



Quantiles

Quantile is a measure that tells how many values in a dataset are above or below a certain limit. It divides the members of the dataset into equally-sized subgroups.

Siddhardhan

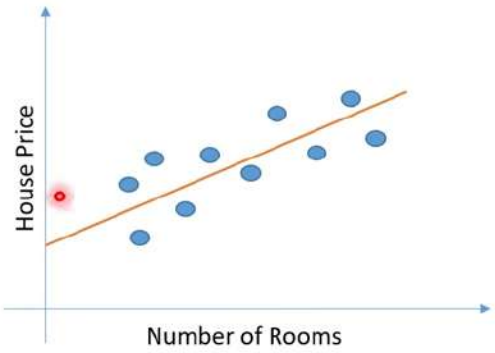
Correlation & Causation

Math for Machine Learning

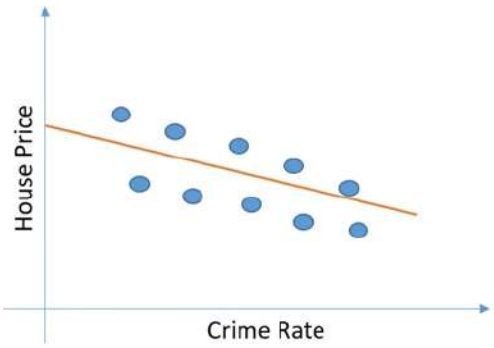
Correlation

Correlation is a measure that determines the extent to which two variables are related to each other in a dataset. But it doesn't mean that one event is the cause of the other event.

Positive Correlation

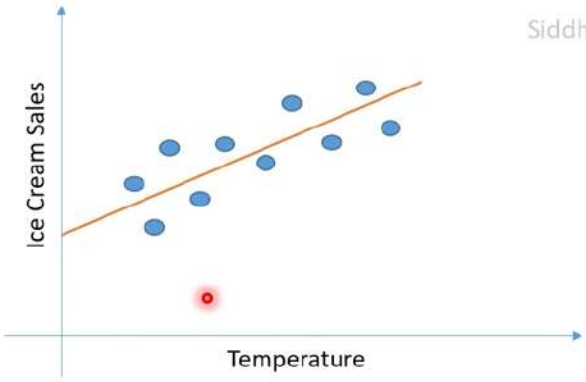


Negative Correlation



Causation

In statistics, Causation means that one event causes another event to occur. Thus, there is a cause and effect relationship between the two variables in a dataset.



Siddhardhan

Hypothesis Testing:
- Null Hypothesis & Alternative Hypothesis

Math for Machine Learning

Hypothesis

Hypothesis is an assumption that is made based on the observations of an experiment.

Hypothesis

Null Hypothesis

Null Hypothesis (H_0) is the commonly accepted fact.

Example: Ptolemy proposed that sun, stars and other planets revolve around the earth.

Alternative Hypothesis

Alternative Hypothesis (H_a) is opposite to null hypothesis and it challenges the null hypothesis.

Example: Aryabhata proposed that earth and other planets revolve around the sun.

Hypothesis Testing

Hypothesis is an assumption that is made based on the observations of an experiment. Hypothesis Testing is a method carried out to tests the assumptions made in the experiment.

Pharmaceutical Company

Drug A

Drug B

Headache

Hypothesis Testing

GROUP 1

Drug A

[12, 8, 13, 10, 7]

(Time taken for recovery in minutes)

Average Time taken = 10 minutes

GROUP 2

Drug B

[15, 12, 18, 16, 14]

(Time taken for recovery in minutes)

Average Time taken = 15 minutes

NULL HYPOTHESIS: Drug A takes 10 minutes on an average to cure headache; Drug B takes 15 minutes on an average to cure headache. Hence, Drug A is more quicker.

Hypothesis Testing

NULL HYPOTHESIS (H_0): Drug A is more quicker than Drug B.

ALTERNATIVE HYPOTHESIS (H_a): Drug B is more quicker than Drug A.

Drug B

Possible Outcomes of Hypothesis Testing:

- Reject the Null Hypothesis
- Fail to reject the Null Hypothesis

Siddhardhan

Probability for Machine Learning

Math for Machine Learning

Probability

Siddhardhan

Probability for Machine Learning

Math for Machine Learning

Probability

What is Probability ?

Probability is a branch of Mathematics that deals with calculating the likelihood of a given event to occur.

Simple Examples:

1. Roll a Dice
2. Toss a coin
3. Bag containing different coloured balls

Topics covered in this module:

1. Basics of Probability
2. Random Variables
3. Probability Distributions
4. Maximum Likelihood
5. Bayes Theorem
6. Information Theory
7. Cross Entropy
8. Information Gain

Basics of Probability

$$\text{Probability of an event to occur} = \frac{\text{Number of ways an event can occur}}{\text{Total number of outcomes}}$$

(H, T)

Possible Outcomes

$$P(H) = \frac{1}{2}$$

$$P(T) = \frac{1}{2}$$

(1, 2, 3, 4, 5, 6)

Possible Outcomes

$$P(5) = \frac{1}{6}$$

$$P(even) = \frac{3}{6}$$

$$P(5) = 0.16$$

Basics of Probability

$$\text{Probability of an event to occur} = \frac{\text{Number of ways an event can occur}}{\text{Total number of outcomes}}$$

Head Tail

(H, T)

Possible Outcomes

$$P(H) = \frac{1}{2}$$

$$P(T) = \frac{1}{2}$$

(1, 2, 3, 4, 5, 6)

Possible Outcomes

$$P(5) = \frac{1}{6}$$

$$P(5) = 0.16$$

$$P(even) = \frac{3}{6}$$

Siddhardhan

Random Variables; Types of Random Variables

Siddhardhan

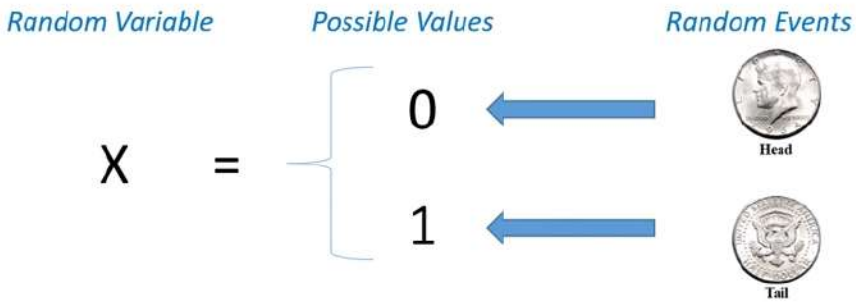
Math for Machine Learning

Random Variables

A Random Variable is a numerical description of the outcomes of Random events.

In other words, a random variable maps the outcomes of random events to numerical values.

Consider Tossing a Coin



Random Variables

Few Examples of Random Variables:

$$X = \begin{cases} 0, & \text{if Heads} \\ 1, & \text{if Tail} \end{cases}$$

Y = Weight of a random person in a class

P (Weight of a random person in a class is less than 60 kg)

P (Y < 60)

Applications:

- Turnover of a company in a given time period.
- Price change of an asset over a given time period

Types of Data

Random Variables

Discrete

A discrete random variable takes only discrete or distinct values.

Examples: Coin toss, Colour of the ball.

Continuous

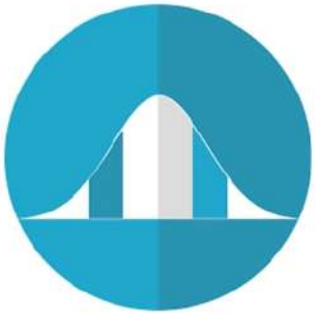
A continuous random variable can take any value in a given range.

Examples: weight of a random person in a class.

Siddhardhan

Probability Distribution for Random Variable

Math for Machine Learning



Probability Distributions

The **probability distribution** for a random variable describes how the probabilities are distributed over the values of the random variable.

Tossing 3 Coins

X = Sum of number of Heads when 3 coins are tossed

HHH = 3	THH = 2
TTT = 0	TTH = 1
HHT = 2	HTT = 1
HTH = 2	THT = 1

Probability Distributions

HHH = 3	THH = 2
TTT = 0	TTH = 1
HHT = 2	HTT = 1
HTH = 2	THT = 1

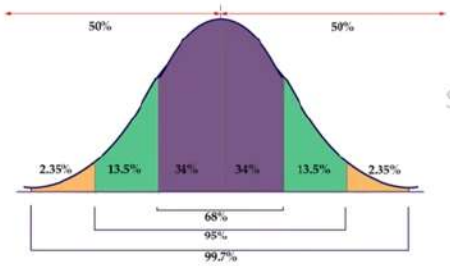
X (No. of Heads)	P(X = x)	P(X = x)
0	1/8	0.125
1	3/8	0.375
2	3/8	0.375
3	1/8	0.125

Discrete Probability Distributions

Siddhardhan

Normal Distribution & Skewness

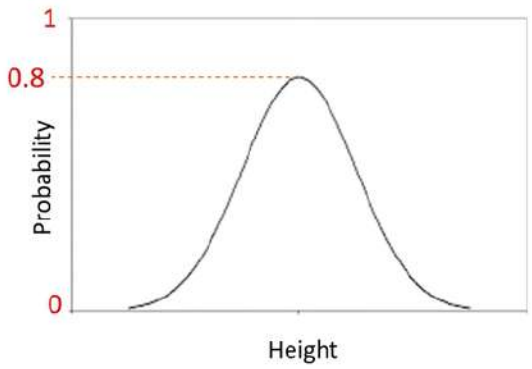
Math for Machine Learning



Normal Distribution

A **normal distribution** is an arrangement of a data set in which most of the data points lie in the middle of the range and the rest taper off symmetrically toward either extreme.

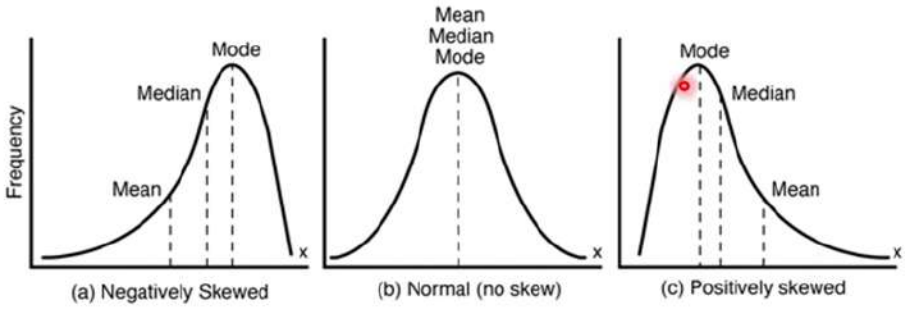
Normal Distribution is also known as **Gaussian Distribution**.



Bell Shaped Curve

Skewness

A data is considered **skewed** when the distribution curve appears distorted or skewed either to the left or to the right, in a statistical distribution.

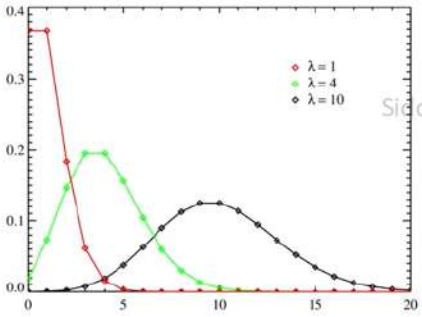


Example: Average income of people in different cities

Siddhardhan

Poisson Distribution

Math for Machine Learning



Poisson Distribution

Poisson Distribution is a probability distribution that measures how many times an event is likely to occur within a specified period of time.

Poisson distribution is used to understand independent events that occur at a constant rate within a given interval of time.

Examples of Poisson Distribution

- Number of accidents occurring in a city from 6 pm to 10 pm
- Number of Patients arriving in an Emergency Room between 10 pm to 12 pm
- How many views does your blog gets in a day

Poisson Distribution

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

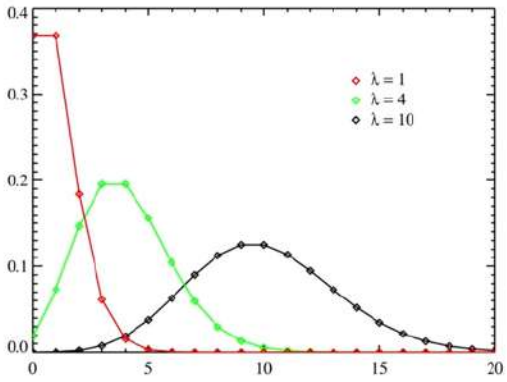
x --> Number of times the event occurs

p(x) --> Probability

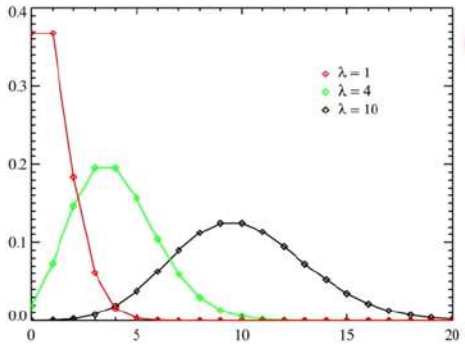
λ --> Mean number of events

x! --> Factorial of x

e --> Euler's Number (2.71828)



Poisson Distribution



Data Science

Siddhardhan