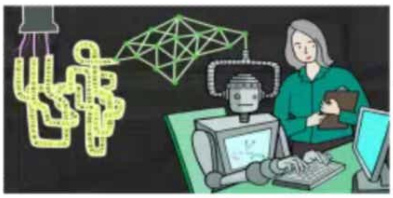
What is Machine Learning ?

- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.



- It is the use of data to help the computer understand various patterns which helps it make better predictions (better accuracy).

What all will you learn in this course?

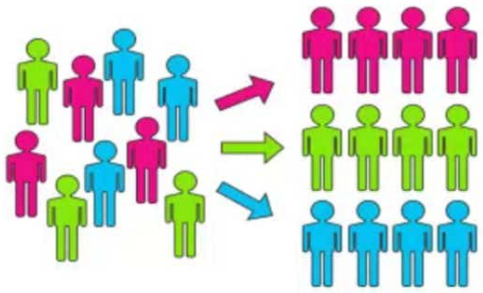
- Introduction to Machine Learning
- Applied Statistics
- Basics of Python
- EDA
- Machine Learning and its Types
- Projects

Why do we use Machine learning ?

- Machine Learning helps us make better predictions. It is like knowing approximately what's going to happen in the future. Almost every field now used machine learning. Let's Discuss a few.
  - **Stock Market Price Predictions (Finance)**
  - **Recommendation System (Entertainment / Sales)**
  - **Customer Segmentation (All Fields)**
  - **Fraud Detection (Insurance)**

Customer Segmentation (All Fields)

- This system helps us segregate the customers on the basis of various factors and then cater to them separately so that we can get the best results.
- We again use various factors to identify and determine which customer falls into which group.



Fraud Detection (Insurance)

- The system predicts if a particular transactions seems to be a fraud or not. This is one of the famous use cases for ML.
- With the advancement in technology, this topic has gain a lot of attention. It is becoming normal transactions are difficult to differentiate from the fraudulent ones as the days pass by.

Getting Started

- There are some basic requirements that is necessary before you dive into the Machine Learning aspect of this course.
  - **You must know how to code. (in any language)**
  - **Basic understanding of statistics (applied statistics)**
  - **Basic understanding of mathematics (applied math)**

Statistics

Descriptive statistics	Probability distributions [Binomial, Poisson & Normal distribution]
Probability & Conditional Probability	Hypothesis Testing

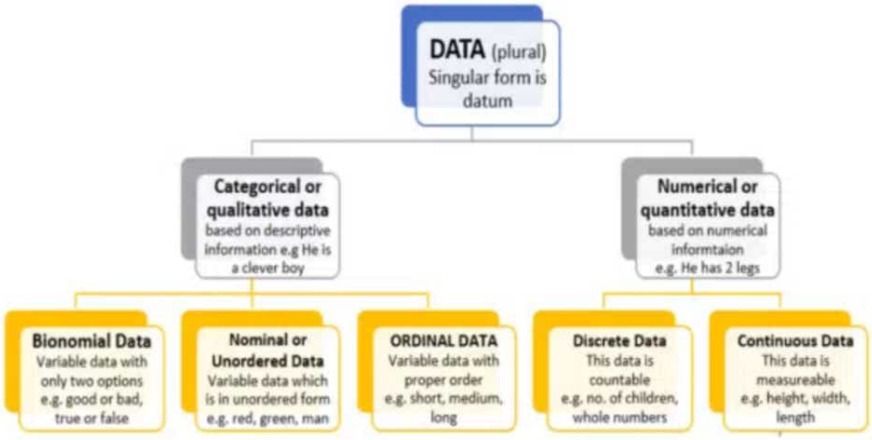
What is Statistics?

- The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.
- It's a branch of science that deals with numeric data. It is mainly used to infer knowledge about big sections (parts / proportions) by trying to understand a smaller part of it.

Why do we need Statistics ?

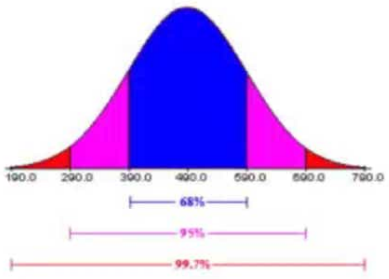
- The major reason is that it helps us understand data better.
- When we understand the data better we work on it better.
- When we work on it better, our deliverables(outcome) after working on the data turns out to be better.

What kind of data do we deal with?



Descriptive Statistics

- Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population.
- Descriptive statistics are broken down into measures of central tendency and measures of variability (spread)



Measures of central tendency

- When we say measures of central tendency, we basically are talking about 3 things the mean the median and the mode.
- The mean, which is also referred as the average or the expected value. It's nothing but average value of the population or the sample. This value gives us an approximate idea of what the entire data looks like.

Mean

- Statistical Formula :
- X = mean
- ∑x = summation of all the values
- N = number of observations
- Simpler formula :
- Mean = Summation of all the values / Total number of observations

$$\overline{X} = \frac{\sum X}{N}$$

Understanding outliers

Person	Income
Mr. X	$1,500
Mr. Y	$1,200
Elon Musk	$1,000,000,000

Mean : $ 333,334,233.33
Approx : $ 333,000,000

Examples (Mean)

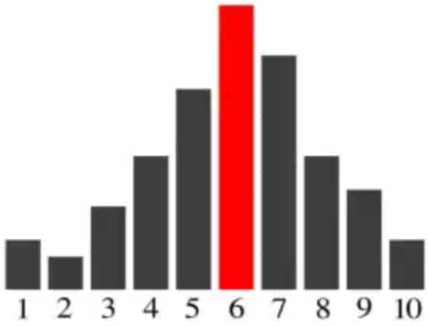
- There are 3 baskets A, B, C having 12, 14, 22 apples respectively from 3 different locations. The farmer Mr. Orchard wants to know how many apples on an average does he have.
  - Answer = 16
- There are 3 baskets A, B, C having 12, 14, 220 apples respectively from 3 different locations. The farmer Mr. Orchard wants to know how many apples on an average does he have.

Example (Median)

- 4 families meet at a restaurant to have dinner together, the manager asks the host for the average number of people in each family so that he can arrange the tables accordingly. The host recollected:
- Family 1 : 3
- Family 2 : 4
- Family 3 : 5
- Family 4 : 5
- Help the host find the median from the above data.
  - Answer : 4.5

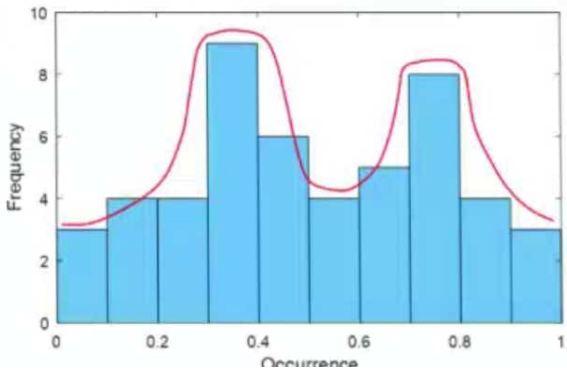
Mode

- Mode is the observation with the highest number of occurrence (frequency) in the data. The data doesn't need to be arranged, we can identify the mode just by glancing over the frequency of each observation.



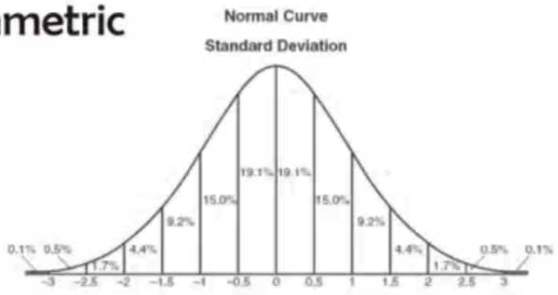
Bimodal

- The data on the right shows that the data has 2 such occurrences that have the high frequencies.
- Example:
  The sale of all the beverages from a shop, the data is most likely to look something like this. The sale of coffee and tea are most likely to outrun the other in that section.



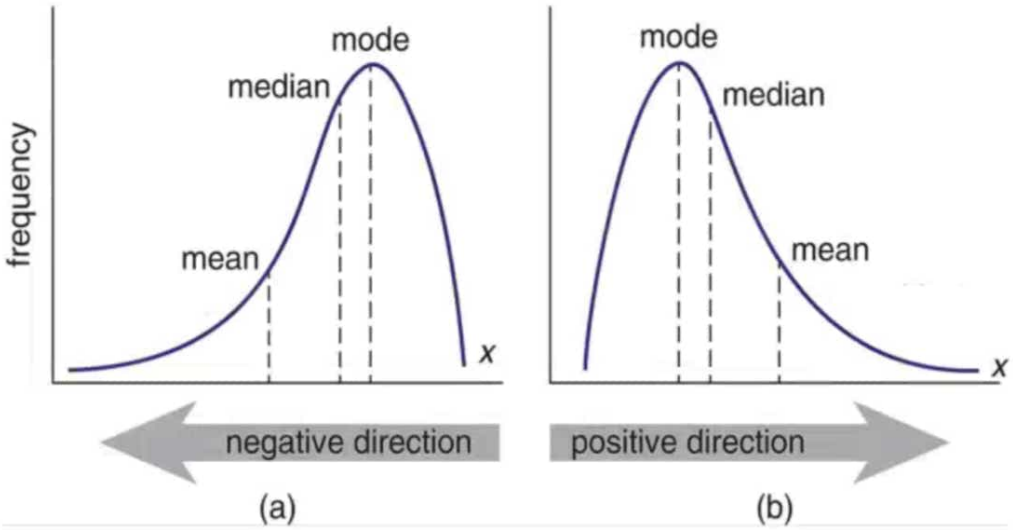
Understanding the bell curve (normal curve)

- The name of the curve is because it's resemblance to a bell.
- The bell shape of the curve portrays the symmetric nature of the curve.



- Example:
  If we pick up a random group of people from the crowd of all age groups, we will find that most of them will be from the in the age of 30-50, a person having an age of 80+ or a kid of age say 10 is really low.

Skewed Data

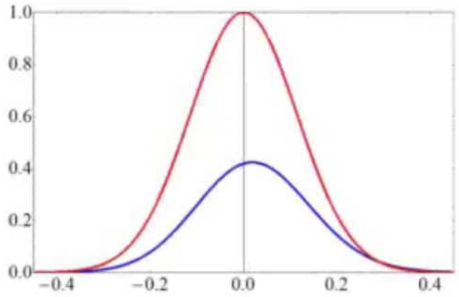


Negatively Skewed:
Mode>Median>Mean

Positively Skewed:
Mean>Median>Mode

What is Measure of Spread

- It is a measure that gives us an idea of how spread across the data is.
- Just because 2 data sets have the same mean, doesn't make the data sets similar.



Examples

- There are 2 classes with 5 student's each. The obtained marks are follows:
  - Class 1 : [-10, 0, 10, 20, 30]
  - Class 2 : [08, 09 10, 11, 12]

Variance (σ2)

Variance Definition in Statistics:

- It is a measure of how data points differ from the mean. According to layman's terms, it is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

Variance Formula

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

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

Examples

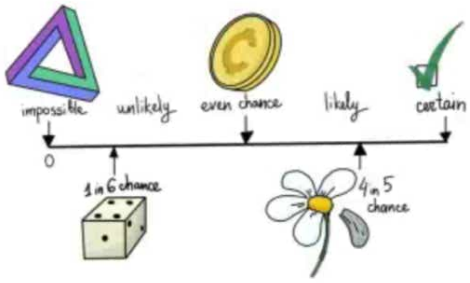
- There are 2 classes with 5 student's each. The obtained marks are follows:
  - Class 1 : [-10, 0, 10, 20, 30]
  - Class 2 : [08, 09 10, 11, 12]
- Variance :
- Class 1 : $\frac{(-10-10)2+(0-10)2+(10-10)2+(20-10)2+(30-10)2}{5} = \frac{1000}{5} = 200$
- Class 2 : $\frac{(08-10)2+(09-10)2+(10-10)2+(11-10)2+(12-10)2}{5} = \frac{10}{5} = 2$

Standard Deviation (σ)

- The standard deviation is nothing but the root over of variance. It is a quantity expressing by how much the members of a group differ from the mean value for the group.

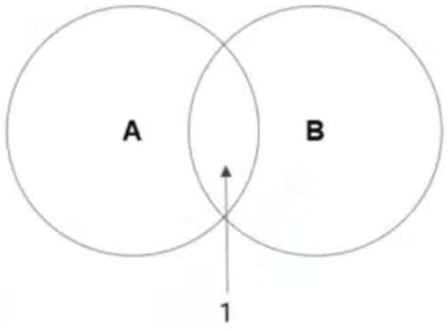
Probability

- The quality or state of being probable; the extent to which something is likely to happen or be the case.
- Basically it tells you the chances of an event to occur or not occur. We can quantify it in terms of a fraction, a decimal or in terms of percentages.
- The value is always between 0 and 1.



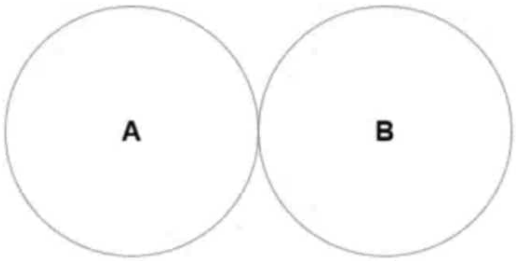
Examples

- There are 52 playing cards in a deck, what is the probability of getting a jack and heart.
- Ans. It's a known fact that there are 4 suits in a deck, and each suit has 1 jack each. So there should be 1/52. Let's test it out.
- P(A) = probability that a card is jack = 4/52 = 1/13
  P(B) = probability that a card is heart = 13/52 = 1/4
- $P(A \cap B) = \frac{1 \times 1}{13 \times 4} = \frac{1}{52}$



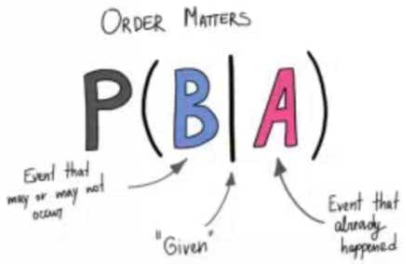
Examples

- What is the probability of getting a 4 and 6 in a roll of a die, considering that it is unbiased.
- A = event where 4 occurs.
  B = event where 6 occurs.
- $P(A \cap B)$ = Where A and B both occur at the same time. (which is not possible)
- Since there is no intersection, therefore
- $P(A \cap B) = 0$



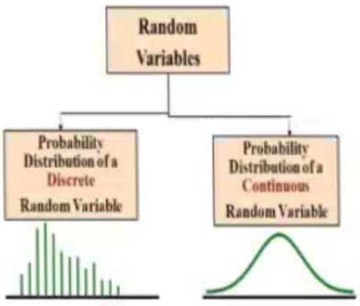
Conditional Probability

- Conditional probability is finding the probability of an event given the other event occurs.
- Formula:
  $P(B|A) = P(B \cap A)/P(A)$
- So basically we try to find the probability of an event based on another event that has already occurred.



Probability Distributions: Discrete and Continuous

- Random Variables play a vital role in probability distributions and also serve as the base for Probability distributions.
- Probability Distribution is a statistical function which links or lists all the possible outcomes a random variable can take, in any random process, with its corresponding probability of occurrence.



Probability Distributions: Discrete and Continuous

- Ex: Probability distribution for a discrete random variable.

No of Heads(X)	Probability P(X)
0	0.25
1	0.5
2	0.25

- If a random variable can take only finite set of values (Discrete Random Variable), then its probability distribution is called as Probability Mass Function or PMF.
- Probability Distribution of discrete random variable is the list of values of different outcomes and their respective probabilities.

Need for Probability Distribution

- According to the definition of random variable, it's the variable which can hold different set of values from the outcome of any random process. However, it lacks the capability to capture the probability of getting those different values. So, probability distribution helps to create a clear picture of all the possible set of values with their respective probability of occurrence in any random process.

- Many decisions in business, insurance, and other real-life situations are made by assigning probabilities to all possible outcomes pertaining to the situation and then evaluating the results
  - Saleswoman can compute probability that she will make 0, 1, 2, or 3 or more sales in a single day. Then, she would be able to compute the average number of sales she makes per week, and if she is working on commission, she will be able to approximate her weekly income over a period of time.
  - An investor wants to compare the risks of two different stock options for his portfolio

Poisson Distribution

- Poisson distribution is a probability distribution that is used to show how many times an event is likely to occur over a specified period.
- In other words, it is a count distribution. Poisson distributions are often used to understand independent events that occur at a constant rate within a given interval of time.

Poisson Probability Distribution

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

λ- mean number of successes over a given interval

$Var(X) = \lambda$

Terminologies in Hypothesis Testing

- Parameter - It is a summary description of a fixed characteristic or measure of the target population. A Parameter denotes the true value that would be obtained if a census rather than a sample were undertaken.
- Ex: Mean (μ), Variance (σ^2), Standard Deviation (σ), Proportion (π)
- Statistic - It is a summary description of a characteristic or measure of the sample. The Sample Statistic is used as an estimate of the population parameter.

Terminologies in Hypothesis Testing

- Parameter - It is a summary description of a fixed characteristic or measure of the target population. A Parameter denotes the true value that would be obtained if a census rather than a sample were undertaken.
- Ex: Mean (μ), Variance (σ^2), Standard Deviation (σ), Proportion (π)
- Statistic - It is a summary description of a characteristic or measure of the sample. The Sample Statistic is used as an estimate of the population parameter.
- Ex: Sample Mean ($\bar{x}$), Sample Variance (S^2), Sample Standard Deviation (S), Sample Proportion (p)

Hypothesis Testing

- **Standard Error (SE)** - It is very similar to the standard deviation. Both are measures of spread.
- The higher the number, the more spread out the data is. The standard error uses statistics (sample data) standard deviation use parameters (population data).
- The standard error tells us how far our sample statistic (like the sample mean) deviates from the actual population mean. The larger our sample size, the smaller the SE.
- In other words, the larger our sample size, the closer our sample mean is to the actual population mean.

Hypothesis Testing

- **Null Hypothesis (H0)** - A statement in which no difference or effect is expected.
- If the null hypothesis is not rejected, no changes will be made.
- The word "null" in this context means that it's a commonly accepted fact that researchers nullify.
- It doesn't mean that the statement is null itself!
- **Alternate Hypothesis (H1 or Ha)** - A statement in which some difference or effect is expected.
- Accepting the alternative hypothesis will lead to changes in opinions or actions. It is the opposite of the null hypothesis.

Hypothesis Testing

Null vs. Alternative Hypothesis

Null Hypothesis

H_0

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include =, ≤, or ≥ sign.

Alternative Hypothesis

H_a

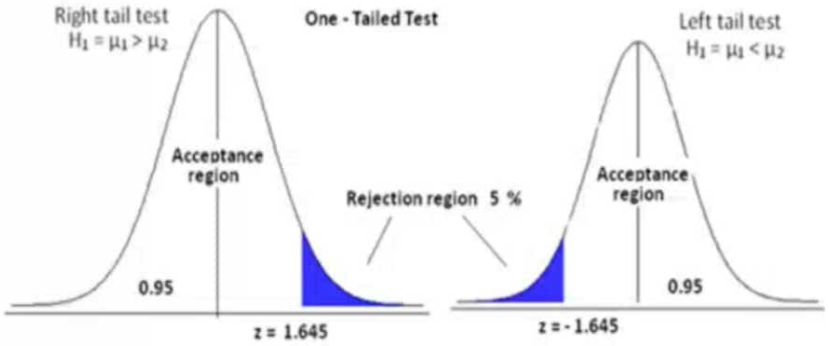
A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include a ≠, >, or < sign.

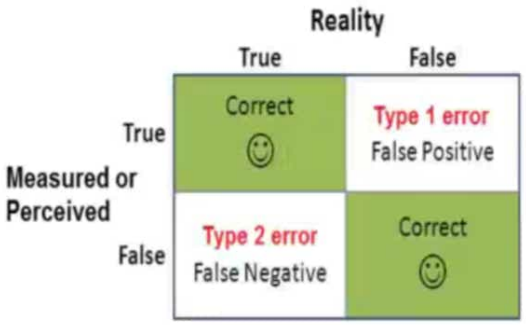
Hypothesis Testing

- **One Tailed Test** - It is a statistical hypothesis test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.



Hypothesis Testing

- **Types of Errors** - There are two types of errors that relate to incorrect conclusions about the null hypothesis.
- **Type I Error** - It occurs when the sample results, lead to the rejection of the null hypothesis when it is in fact true. Type-I errors are equivalent to false positives.
- **Type II Error** - It occurs when based on the sample results, the null hypothesis is not rejected when it is in fact false. Type-II errors are equivalent to false negatives.



Hypothesis Testing

- **Level of Significance (α)** - It is the probability of making a Type-I error and it is denoted by alpha (α). Alpha is the maximum probability that we have a Type-I error. For a 95% confidence level, the value of alpha is 0.05. This means that there is a 5% probability that we will reject a true null hypothesis.
- **P-Value** - A p-value is used in hypothesis testing to help us support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should

Importing the Data

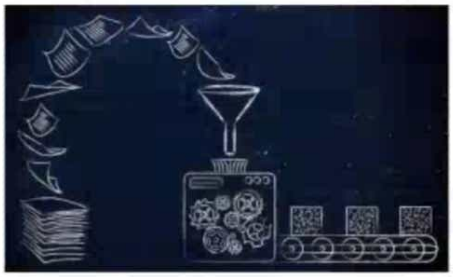
- The data can be in various formats and distributed amongst various files. Combining them is one of the task that is crucial before processing it any further.
- Many a times the data is unstructured, understanding the flow, pattern and converting it into a consistent format is another task that comes under this process.
- Data Cleaning : Sometimes there are missing values in the data that need to be taken care of, there are various of doing so.

Exploratory Data Analysis (EDA)

- Viewing the data in various ways to understand the structure.
- Understanding how the values distributed in the columns
- Use Visual and Non-Visual Methods to understand the data.
- Make a note of all the inferences, insights and assumptions that you gain or make for the visuals.

Data Transformation

- If there is any discrepancy in the data type of the column values or you need to modify the data into a consistent format.
- Sometimes you need to make different column values from the existing ones or remove a few columns as they don't add any information that helps us build a better model.



Model selection

- After understanding how the data is and transforming it in the best way possible for further computation. Choosing the model is an important task. Just because a complex model gives us better accuracy (generally) we don't start of with them.
- We start of with the weak models, and gradually move up the ladder. Each model is selected trained, tested and tuned till we feel that that's the best that the model can do on this data set.

Deployment

- After being satisfied with the results we deploy the model in the real world to see how it works on the data from the real world.
- The task doesn't end here. If the model works fine we try to make it work better. If the model fails to work, then we take it up again and go through the entire cycle all over again. It's a continuous process.

Types of machine learning algorithms