Week 4
Artificial Intelligence and Machine Learning

Probability

- Basic concepts

Probability, loosely speaking, concerns the study of uncertainty. Probability can be thought of as the fraction of times an event occurs, or as a degree of belief about an event. We then would like to use this probability to measure the chance of something occurring in an experiment. We often quantify uncertainty in the data, uncertainty in the machine learning model, and uncertainty in the predictions produced by the model.

Probability is a measure of uncertainty..
The **probability** of an event refers to the likelihood that the event will occur. Mathematically, the probability that an event will occur is expressed as a number between 0 and 1. Notionally, the probability of event A is represented by P(A).

- If P(A) equals 0, 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 1, event A will almost definitely occur.

In a statistical experiment, the sum of probabilities for all possible outcomes is equal to one. This means, for example, that if an experiment can have three possible outcomes (A, B, and C), then P(A) + P(B) + P(C) = 1.

Fundamental Concepts in Probability

Definitions

- random experiment = experiment (action) whose result is uncertain (cannot be predicted with certainty) before it is performed
- trial = single performance of the random experiment
- outcome = result of a trial
- sample space = the set of all possible outcomes of a random experiment
- event = subset of the sample space (to which a probability can be assigned) = a collection of possible outcomes
- sure event = sample space (an event for sure to occur)
- impossible event = empty set (an event impossible to occur)

Sample Spaces and Events

Rolling an ordinary six-sided die is a familiar example of a *random experiment*, an action for which all possible outcomes can be listed, but for which the actual outcome on any given trial of the experiment cannot be predicted with certainty. In such a situation we wish to assign to each outcome, such as rolling a two, a number, called the *probability* of the outcome, that indicates how likely it is that the outcome will occur. Similarly, we would like to assign a probability to any *event*, or collection of outcomes, such as rolling an even number, which indicates how likely it is that the event will occur if the experiment is performed.

Definition: random experiment

A *random experiment* is a mechanism that produces a definite outcome that cannot be predicted with certainty. The sample space associated with a random experiment is the set of all possible outcomes. An event is a subset of the sample space.

Definition: Element and Occurrence

An event E is said to occur on a particular trial of the experiment if the outcome observed is an element of the set E.

Example

Sample Space for a single coin

Construct a sample space for the experiment that consists of tossing a single coin.

Solution

The outcomes could be labeled h for heads and t for tails. Then the

sample

space is the set: $S=\{h,t\}$

Example

Sample Space for a single die

Construct a sample space for the experiment that consists of rolling a single die. Find the events that correspond to the phrases "an even number is rolled" and "a number greater than two is rolled."

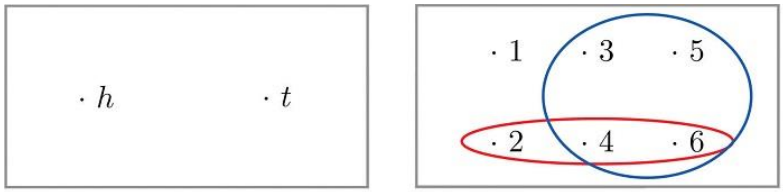
Solution:

The outcomes could be labeled according to the number of dots on the top face of the die. Then the sample space is the set $S=\{1,2,3,4,5,6\}$The outcomes that are even are $2,4,\text{and}6$, so the event that corresponds to the phrase "an even number is rolled" is the set $\{2,4,6\}$, which it is natural to denote by the letter E. We write $E=\{2,4,6\}$.

Similarly the event that corresponds to the phrase "a number greater than two is rolled" is the set $T=\{3,4,5,6\}$, which we have denoted T.

A graphical representation of a

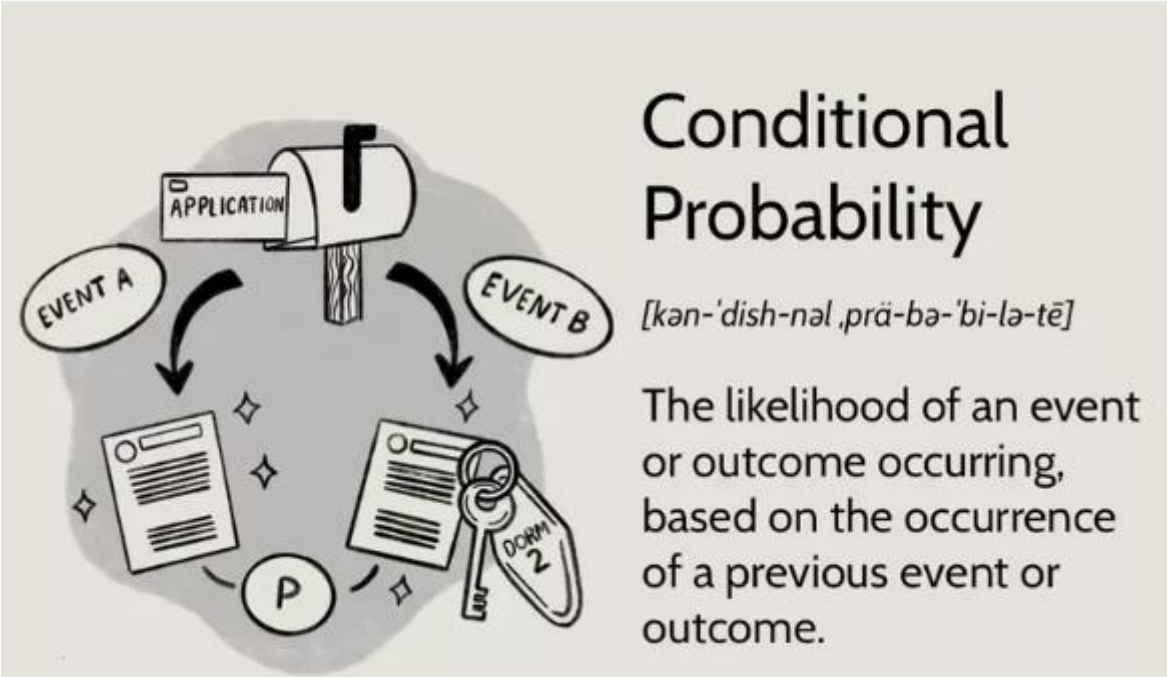
sample space and events is a *Venn diagram*, as shown in Figure 3.1.13.1.1. In general the sample space S is represented by a rectangle, outcomes by points within the rectangle, and events by ovals that enclose the outcomes that compose them.



(a): One Coin Toss (b): One Die Roll

Figure 3.1.13.1.1: Venn Diagrams for Two Sample Spaces

Conditional Probability:



Conditional Probability:
Condition Probability is the probability of an event occurring given that another event has already occurred.
i.e. we try to calculate the probability of the second event given that the first event has already occurred.
Example:
It will be raining at the end of the hottest day.
Here the probability of occurrence of rainfall depends on the temperature throughout the day.
Mathematical Formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where,

$P(A|B)$: $Probability\ of\ event\ A\ given\ that\ event\ B\ happened$

$P(A \cap B)$: $Probability\ that\ both\ the\ events\ A\ and\ B\ occur$

$P(B)$: $Probability\ of\ event\ B$

Note: $P(A|B) \neq P(B|A)$

Let's understand the conditional probability with an example

Example:
Let there are 100 (47 boys, 53 girls) students in a classroom, the given table describes the number of students (boys and girls) who pass or fail in the exam.

	Pass	Fail	Total
Boys	**15**	**32**	**47**
Girls	**29**	**24**	**53**
Total	**44**	**56**	**100**

Let A: represent passed in the exam

And B: represent student is a Boy

- Probability of a Boy to Pass the exam:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{15}{47}$$

- Probability that the student passed is a boy

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{15}{44}$$

What is a Joint Probability?

A joint probability, in probability theory, refers to the probability that two events will both occur. In other words, joint probability is the likelihood of two events occurring together.

Formula for Joint Probability

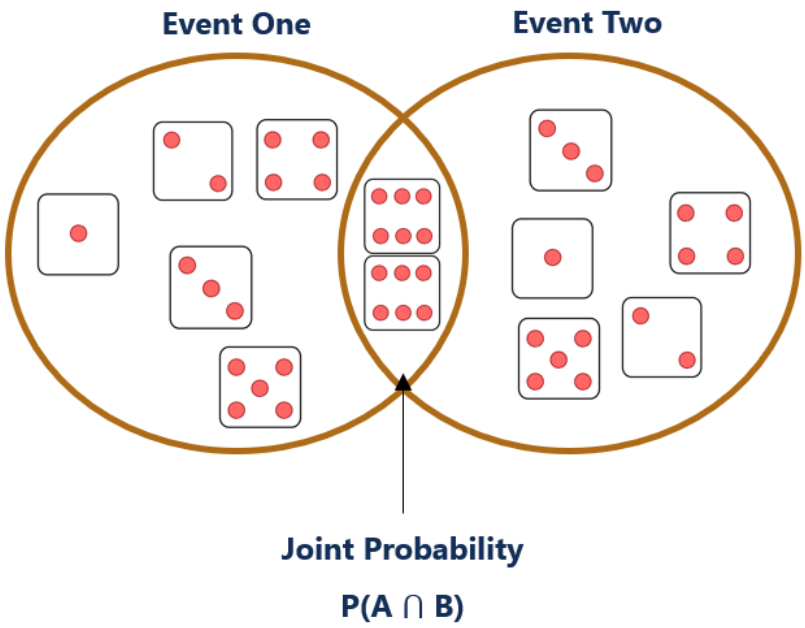
$$\textbf{Joint Probability} = P(A \cap B) = P(A) \times P(B)$$

Where:

- **P(A ∩ B)** is the notation for the joint probability of event "A" and "B".
- **P(A)** is the probability of event "A" occurring.

- **P(B)** is the probability of event "B" occurring.

  A joint probability can be visually represented through a Venn diagram. Consider the joint probability of rolling two 6's in a fair six-sided dice:



- Shown on the Venn diagram above, the joint probability is where both circles overlap each other. It is called the "intersection of two events."

Example 1
What is the joint probability of rolling the number five twice in a fair six-sided dice?
Event "A" = The probability of rolling a 5 in the first roll is 1/6 = 0.1666.
Event "B" = The probability of rolling a 5 in the second roll is 1/6 = 0.1666.
Therefore, the joint probability of event "A" and "B" is P(1/6) x P(1/6) = 0.02777 = **2.8%**.
Example 2
What is the joint probability of getting a head followed by a tail in a coin toss?
Event "A" = The probability of getting a head in the first coin toss is 1/2 = 0.5.
Event "B" = The probability of getting a tail in the second coin toss is 1/2 = 0.5.
Therefore, the joint probability of event "A" and "B" is P(1/2) x P(1/2) = 0.25 = **25%**.

Bayes' Theorem:

Bayes' Theorem or Bayes' Rule is named after Reverend Thomas Bayes. It describes the probability of an event, based on prior knowledge of conditions that might be related to that event.

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

For example: There are 3 bags, each containing some white marbles and some black marbles in each bag. If a white marble is drawn at random. With probability to find that this white marble is from the first bag. In cases like such, we use the Bayes' Theorem.

Bayes' theorem is an extension of Conditional Probability.
It includes two conditional probabilities.
It gives the relation between conditional probability and its reverse form.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

where,

$P(A)$: $Marginal\ probability\ of\ A$

$P(B)$: $MArginal\ probability\ of\ B$

$P(A|B)$: $Conditional\ probability\ of\ A\ given\ B$

$P(B|A)$: $Conditional\ probability\ of\ B\ given\ A$

Probability Distributions

Probability can be used for more than calculating the likelihood of one event; it can summarize the likelihood of all possible outcomes.
A thing of interest in probability is called a random variable, and the relationship between each possible outcome for a random variable and their probabilities is called a probability distribution.

Probability distributions are an important foundational concept in probability and the names and shapes of common probability distributions will be familiar. The structure and type of the probability distribution varies based on the properties of the random variable, such as continuous or discrete, and this, in turn, impacts how the distribution might be summarized or how to calculate the most likely outcome and its probability.

Discrete Probability Distributions

A discrete probability distribution summarizes the probabilities for a discrete random variable.
The probability mass function, or PMF, defines the probability distribution for a discrete random variable. It is a function that assigns a probability for specific discrete values.
A discrete probability distribution has a cumulative distribution function, or CDF. This is a function that assigns a probability that a discrete random variable will have a value of less than or equal to a specific discrete value.
Probability Mass Function. Probability for a value for a discrete random variable.
Cumulative Distribution Function. Probability less than or equal to a value for a random variable.
The values of the random variable may or may not be ordinal, meaning they may or may not be ordered on a number line, e.g. counts can, car color cannot. In this case, the structure of the PMF and CDF may be discontinuous, or may not form a neat or clean transition in relative probabilities across values.
The expected value for a discrete random variable can be calculated from a sample using the mode, e.g. finding the most common value. The sum of probabilities in the PMF equals to one.
Some examples of well known discrete probability distributions include:
Poisson distribution.
Bernoulli and binomial distributions.
Multinoulli and multinomial distributions.
Discrete uniform distribution.
Some examples of common domains with well-known discrete probability distributions include:
The probabilities of dice rolls form a discrete uniform distribution.
The probabilities of coin flips form a Bernoulli distribution.
The probabilities car colors form a multinomial distribution.

Discrete Random Variables in Probability distribution

A discrete random variable can only take a finite number of values. To further understand this, let's see some examples of discrete random variables:

1. X = {sum of the outcomes when two dice are rolled}. Here, X can only take values like {2, 3, 4, 5, 6….10, 11, 12}.
2. X = {Number of Heads in 100 coin tosses}. Here, X can take only integer values from [0,100].

- **Discrete Probability Distribution:** The one with a limited number of probability values corresponding to discrete random variables.

  Ex.: Occurrence of the number of heads when a fair coin is tossed thrice.

X	0	1	2	3
$p(X)$	1/8	3/8	3/8	1/8

Continuous Probability Distributions

A continuous probability distribution summarizes the probability for a continuous random variable.

The probability distribution function, or PDF, defines the probability distribution for a continuous random variable. Note the difference in the name from the discrete random variable that has a probability mass function, or PMF.

Like a discrete probability distribution, the continuous probability distribution also has a cumulative distribution function, or CDF, that defines the probability of a value less than or equal to a specific numerical value from the domain.

- **Probability Distribution Function**. Probability for a value for a continuous random variable.
- **Cumulative Distribution Function**. Probability less than or equal to a value for a random variable.

As a continuous function, the structure forms a smooth curve.

Some examples of well-known continuous probability distributions include:

- Normal or Gaussian distribution.
- Power-law distribution.
- Pareto distribution.

Some examples of domains with well-known continuous probability distributions include:

- The probabilities of the heights of humans form a Normal distribution.
- The probabilities of movies being a hit form a Power-law distribution.
- The probabilities of income levels form a Pareto distribution.

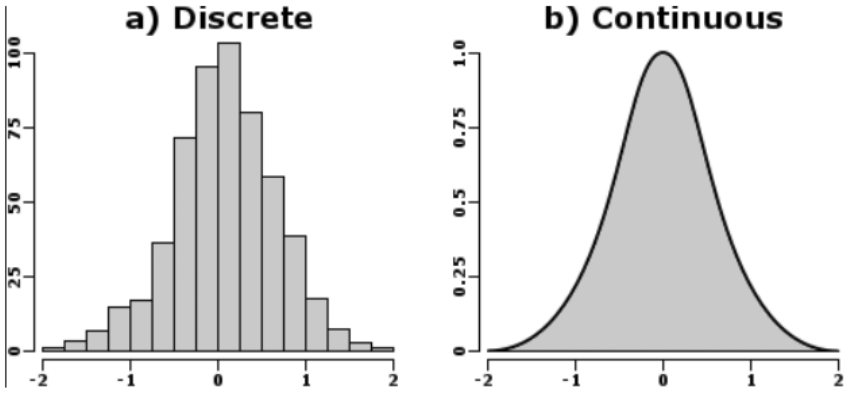
Continuous Random Variable in Probability distribution

A continuous random variable can take infinite values in a continuous domain. Let's see an example of a dart game.

Suppose, we have a dart game in which we throw a dart where the dart can fall anywhere between [-1,1] on the x-axis. So if we define our random variable as the x-coordinate of the position of the dart, X can take any value from [-1,1]. There are infinitely many possible values that X can take. (X = {0.1, 0.001, 0.01, 1,2, 2.112121 …. and so on}.

- **Continuous Probability Distribution:** In this distribution, the variable can take any values in the given interval.

  Ex.: Probability of picking a real number between 0 and 1.



Univariate Analysis

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

What Is the Central Limit Theorem (CLT)?

In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a "bell curve") as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

Put another way, CLT is a statistical premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population. Furthermore, these samples approximate a normal distribution, with their variances being approximately equal to the variance of the population as the sample size gets larger, according to the law of large numbers.

Although this concept was first developed by Abraham de Moivre in 1733, it was not formalized until 1930, when noted Hungarian mathematician George Pólya dubbed it the central limit theorem.12

KEY TAKEAWAYS

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
- Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
- A sufficiently large sample size can predict the characteristics of a population more accurately.
- CLT is useful in finance when analyzing a large collection of securities to estimate portfolio distributions and traits for returns, risk, and correlation.

The Central Limit Theorem states that as the sample size grows higher, the sample size of the sampling values approaches a normal distribution, regardless of the form of the data distribution. The mean of sample means will be the population mean, according to the **Central Limit Theorem**.

Likewise, if you average all the degrees of separation in your sample, you'll get the population's true standard deviation.

The sample mean is equal to the population mean.

The sample standard deviation is equal to the population standard deviation divided by the square root of the sample size.

This theorem is usually employed in cases where the size of the distribution is considerable, preferably more than 30.

Hypothesis Testing in Machine Learning

The process of hypothesis testing is to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample. The same inferences are drawn for different machine learning models through T-test which I will discuss in this tutorial.

For drawing some inferences, we have to make some assumptions that lead to two terms that are used in the hypothesis testing.

- Null hypothesis: It is regarding the assumption that there is no anomaly pattern or believing according to the assumption made.
- Alternate hypothesis: Contrary to the null hypothesis, it shows that observation is the result of real effect.

P value

It can also be said as evidence or level of significance for the null hypothesis or in machine learning algorithms. It's the significance of the predictors towards the target.

Generally, we select the level of significance by 5 %, but it is also a topic of discussion for some cases. If you have a strong prior knowledge about your data functionality, you can decide the level of significance.

On the contrary of that if the p-value is less than 0.05 in a machine learning model against an independent variable, then the variable is considered which means there is heterogeneous behavior with the target which is useful and can be learned by the machine learning algorithms.

The steps involved in the hypothesis testing are as follow:

- Assume a null hypothesis, usually in machine learning algorithms we consider that there is no anomaly between the target and independent variable.
- Collect a sample
- Calculate test statistics
- Decide either to accept or reject the null hypothesis

Calculating test or T statistics

For Calculating T statistics, we create a scenario.

Suppose there is a shipping container making company which claims that each container is 1000 kg in weight not less, not more. Well, such claims look shady, so we proceed with gathering data and creating a sample.
After gathering a sample of 30 containers, we found that the average weight of the container is 990 kg and showing a standard deviation of 12.5 kg.
So calculating test statistics:
T = (Mean - Claim)/ (Standard deviation / Sample Size^(1/2))
Which is -4.3818 after putting all the numbers.
Now we calculate t value for 0.05 significance and degree of freedom.
Note: `Degree of Freedom = Sample Size - 1`
`From T table the value will be -1.699.`

After comparison, we can see that the generated statistics are less than the statistics of the desired level of significance. So we can reject the claim made.

You can calculate the t value using `stats.t.ppf()` function of `stats` class of `scipy` library.

Errors

As hypothesis testing is done on a sample of data rather than the entire population due to the unavailability of the resources in terms of data. Due to inferences are drawn on sample data the hypothesis testing can lead to errors, which can be classified into two parts:

- Type I Error: In this error, we reject the null hypothesis when it is true.
- Type II Error: In this error, we accept the null hypothesis when it is false.

Other Approaches

A lot of different approaches are present to hypothesis testing of two models like creating two models on the features available with us. One model comprises all the features and the other with one less. So we can test the significance of individual features. However feature inter-dependency affect such simple methods.

In regression problems, we generally follow the rule of P value, the feature which violates the significance level are removed, thus iteratively improving the model.

Different approaches are present for each algorithm to test the hypothesis on different features.

Understanding P-values | Definition and Examples

The ***p* value** is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.
P values are used in hypothesis testing to help decide whether to reject the null hypothesis. The smaller the *p* value, the more likely you are to reject the null hypothesis.

What exactly is a *p* value?

The ***p* value**, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your **test statistic**, which is the number calculated by a statistical test using your data.

The *p* value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The *p* value gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis.

The *p* value is a proportion: if your *p* value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true.

Example: Test statistic and *p* valueIf the mice live equally long on either diet, then the test statistic from your *t* test will closely match the test statistic from the null hypothesis (that there is no difference between groups), and the resulting *p* value will be close to 1. It likely won't reach exactly 1, because in real life the groups will probably not be perfectly equal.

If, however, there is an average difference in longevity between the two groups, then your test statistic will move further away from the values predicted by the null hypothesis, and the *p* value will get smaller. The *p* value will never reach zero, because there's always a possibility, even if extremely unlikely, that the patterns in your data occurred by chance.

Interpreting test statistics

For any combination of sample sizes and number of predictor variables, a statistical test will produce a predicted distribution for the test statistic. This shows the most likely range of values that will occur if your data follows the null hypothesis of the statistical test.

The more extreme your test statistic – the further to the edge of the range of predicted test values it is – the less likely it is that your data could have been generated under the null hypothesis of that statistical test.

The agreement between your calculated test statistic and the predicted values is described by the ***p* value**. The smaller the *p* value, the less likely your test statistic is to have occurred under the null hypothesis of the statistical test.

Because the test statistic is generated from your observed data, this ultimately means that the smaller the *p* value, the less likely it is that your data could have occurred if the null hypothesis was true.

Test statistic exampleYour calculated *t* value of 2.36 is far from the expected range of *t* values under the null hypothesis, and the *p* value is < 0.01. This means that you would expect to see a *t* value as large or larger than 2.36 less than 1% of the time if the true relationship between temperature and flowering dates was 0.

Therefore, it is statistically unlikely that your observed data could have occurred under the null hypothesis. Using a significance threshold of 0.05, you can say that the result is **statistically significant**.

Statistics:

Statistics are **used to summarize and make inferences about a large number of data points.** In Data Science and Machine Learning, you will often come across the following terminology

- Centrality measures
- Distributions (especially normal)

Centrality measures and measures of spreads

Mean:

Mean is just an **average of numbers**. To find out mean, you have to sum the numbers and divide it with the number of numbers. For example, the mean of [1,2,3,4,5] is 15/5 = 3.

$$Mean = 1/N * \sum_{n=1}^{N} x_n$$

Median:

Median is the **middle element of a set of numbers** when they are arranged in ascending order. For example, numbers [1,2,4,3,5] are arranged in an ascending order [1,2,3,4,5]. The middle one of these is 3. Therefore the median is 3. But what if the number of numbers is even and therefore has no middle number? In that case, you take the average of the two middle-most numbers. For a sequence of 2n numbers in ascending order, average the nth and $(n+1)^{th}$ number to get the median. Example – [1,2,3,4,5,6] has the median (3+4)/2 = 3.5

Mode:

Mode is simply the **most frequent number in a set of numbers**. For example, mode of [1,2,3,3,4,5,5,5] is 5.

Variance:

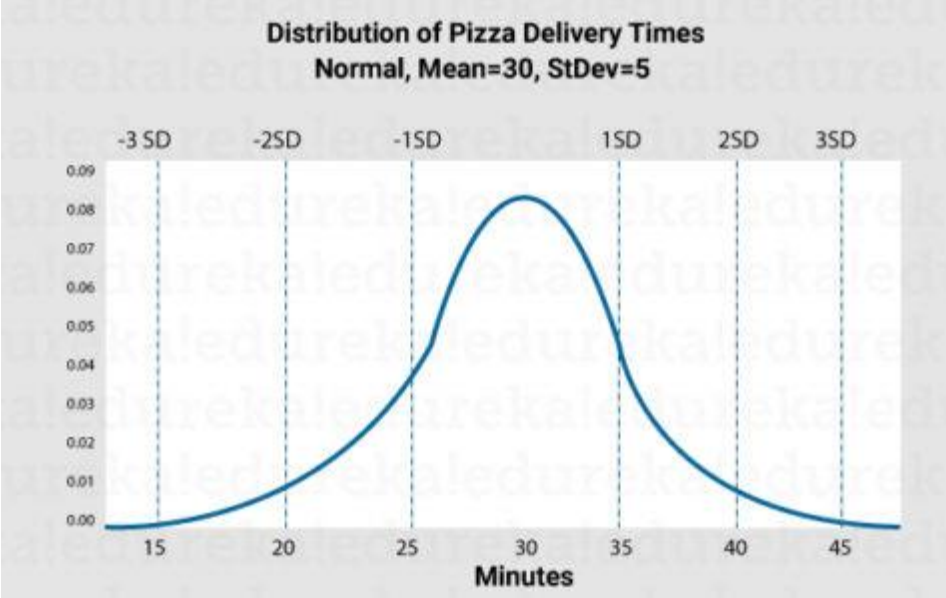
Variance is not a centrality measure. It measures **how your data is spread around the mean**. It is quantified as

$$\sigma^2 = 1/N * \sum_{n=1}^{N} (x_n - \underline{x})^2$$

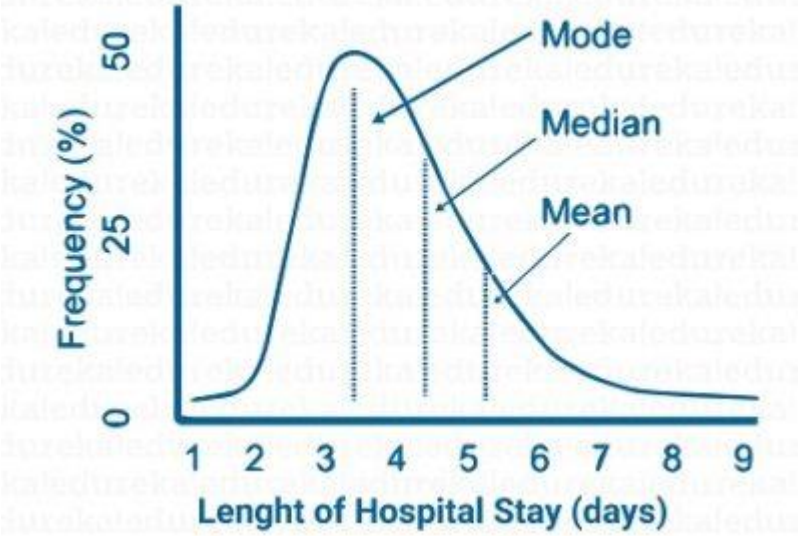
x is the mean of N numbers. You take a point, subtract the mean, take the square of this difference. Do this for all N numbers and average them. The square root of the variance is called the standard deviation. Next, in this article on statistics for machine learning, let us understand Normal Distribution.

Normal Distribution

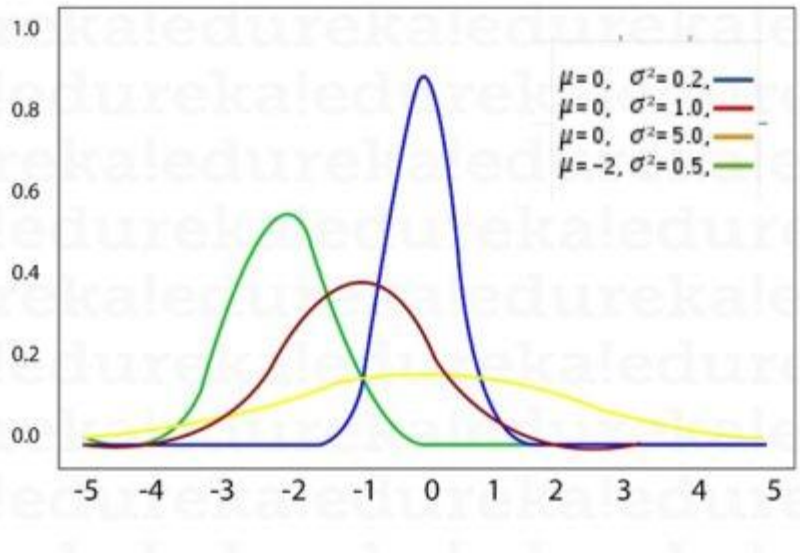
Distribution helps us **understand how our data is spread**. For example, in a sample of ages, we may have young people more than older adults and hence smaller values of age more than greater values. But how do we define a distribution? Consider the example below



The y-axis represents the density. The mode of this distribution is 30 since it is the peak and hence most frequent. We can also locate the median. Median lies at the point on the x-axis where half of the area under the curve is covered. The area under any normal distribution is 1 because the sum of probabilities of all events is 1. For example,



Median in the above case is around 4. This means that area under the curve before 4 is the same as that after 4. Consider another example



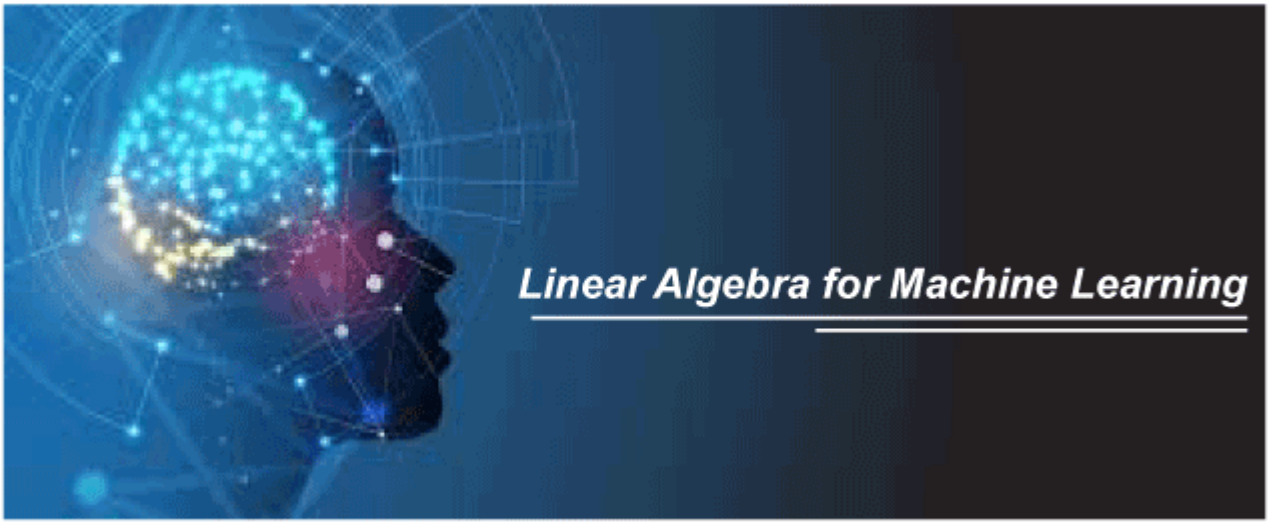
We see three normal distributions. The blue and red ones have the same mean. The red one has a greater variance. Hence, it is more spread out than the blue one. But since the area has to be 1, the peak of the red curve is shorter than the blue curve, to keep the area constant.

Hope you understood the basic statistics and normal distributions. Now, next in this article on statistics for machine learning, let us learn about Linear Algebra.

Linear Algebra for Machine learning

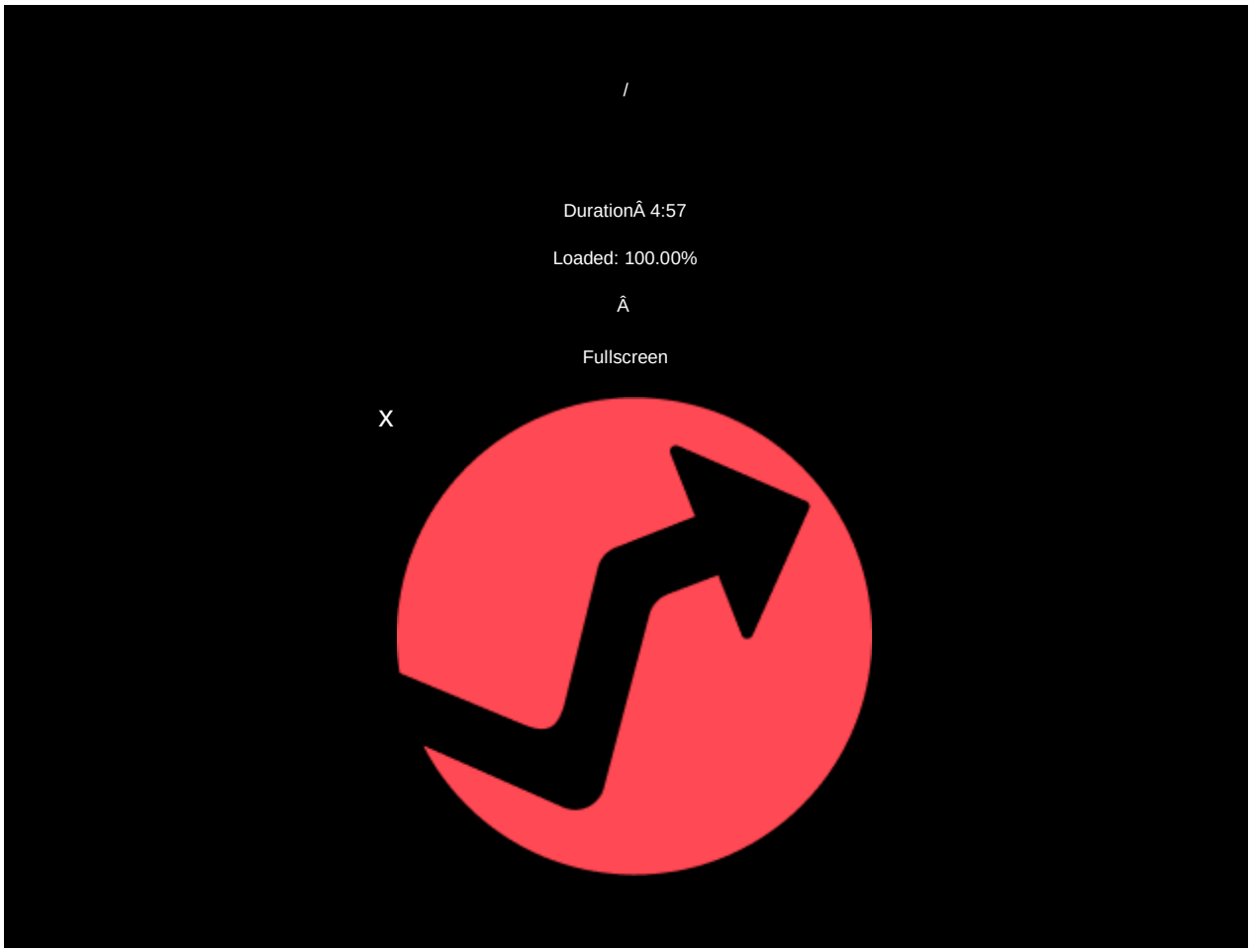
Machine learning has a strong connection with mathematics. Each machine learning algorithm is based on the concepts of mathematics & also with the help of mathematics, one can choose the correct algorithm by considering training time, complexity, number of features, etc. ***Linear Algebra is an essential field of mathematics, which defines the study of vectors, matrices, planes, mapping, and lines required for linear transformation.***

The term Linear Algebra was initially introduced in the early 18th century to find out the unknowns in Linear equations and solve the equation easily; hence it is an important branch of mathematics that helps study data. Also, no one can deny that Linear Algebra is undoubtedly the important and primary thing to process the applications of Machine Learning. It is also a prerequisite to start learning Machine Learning and data science.



Linear algebra plays a vital role and key foundation in machine learning, and it enables ML algorithms to run on a huge number of datasets.

The concepts of linear algebra are widely used in developing algorithms in machine learning. Although it is used almost in each concept of Machine learning, specifically, it can perform the following task:



- Optimization of data.
- Applicable in loss functions, regularisation, covariance matrices, Singular Value Decomposition (SVD), Matrix Operations, and support vector machine classification.
- Implementation of Linear Regression in Machine Learning.

Besides the above uses, linear algebra is also used in neural networks and the data science field.

Basic mathematics principles and concepts like Linear algebra are the foundation of Machine Learning and Deep Learning systems. To learn and understand Machine Learning or Data Science, one needs to be familiar with linear algebra and optimization theory. In this topic, we will explain all the Linear algebra concepts required for machine learning.

Note: Although linear algebra is a must-know part of mathematics for machine learning, it is not required to get intimate in this. It means it is not required to be an expert in linear algebra; instead, only good knowledge of these concepts is more than enough for machine learning.

Why learn Linear Algebra before learning Machine Learning?

Linear Algebra is just similar to the flour of bakery in Machine Learning. As the cake is based on flour similarly, every Machine Learning Model is also based on Linear Algebra. Further, the cake also needs more ingredients like egg, sugar, cream, soda. Similarly, Machine Learning also requires more concepts as vector calculus, probability, and optimization theory. So, we can say that Machine Learning creates a useful model with the help of the above-mentioned mathematical concepts.

Below are some benefits of learning Linear Algebra before Machine learning:

- **Better Graphic experience**
- **Improved Statistics**
- **Creating better Machine Learning algorithms**
- **Estimating the forecast of Machine Learning**
- **Easy to Learn**