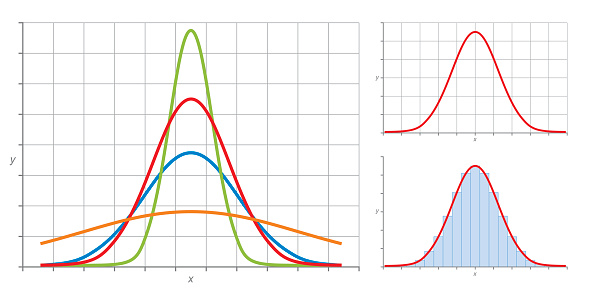
**NORMAL DISTRIBUTION**

In the world of machine learning and data science, probability distribution plays a particularly important role and the core of the probability distribution concept is focused on **Normal Distributions**.

Here we aim to explain the concept of Normal Distribution in the simplest possible way for easy and efficient understanding as well as explain the concept with the help of an example.

**WHAT IS NORMAL DISTRIBUTION?**

If you ask enough people about their shoe size or height you will come up with a graph that is shaped like a **bell** and this is the **Normal Distribution curve**, also known as the **Gaussian distribution** (named after the genius Carl Friedrich Gauss). It is a **symmetric distribution** of **probability over the mean,** indicating that data near the mean occur more often than the data far away from the mean.



NORMAL DISTRIBUTION (BELL SHAPED CURVE)

Now, to understand Normal Distribution it is firstly important for us to know the concept of probability distribution since it forms the base for Normal Distribution.

**WHAT IS PROBABILITY DISTRIBUTION?**

A **probability distribution** is a function that describes the **likelihood** of obtaining the possible values that a **random variable** can assume.

Let us break this down into simpler terms. Suppose we want to predict a variable accurately then we must do the following:

1. First, we need to understand the underlying **behaviour** of our target variable. This means we need to determine the possible outcomes of our target variable and check if the underlying outcomes are [discrete/distinct values or continuous/infinite values](https://www.youtube.com/watch?v=MIPazZSnPFI). Let us take the example of a die. In this the first step would be to take up any value from 1 to 6 (discrete values).
2. Step two would include **assigning probabilities** to the events. The higher the probability, the more likely it is for the event to occur.

Now, if we can start repeating this experiment for a large number of times and start noting the values we retrieve for the variable then we can **group all the values into categories** and for each category, we can start recording the number of times the variable had the value of the category.

For example, we can throw a dice 10000 times and can create 6 categories as there are 6 possible values that a dice can take. After every throw, we can **record** the number of occurrences for each value and note it in the respective category.

We can then **plot a line-chart** where the values of the categories would be the x-values, and the values of the y-axis would represent the category value frequency. We will note that it is the shape of a curve. This curve is known as the **curve of the probability distribution**, and the probability of obtaining the target variable value is known as the variable's probability distribution.

Once we understand how the values are distributed then we can start estimating the **probabilities of the events** (even by using formulas). This results in better understanding of the behaviour of our target variables. The probability distribution’s key parameters are the moments of the sample such as **mean** (average of the sample)**, standard deviation** (shows the variation or dispersion there is from the average)**, skewness** (measures the symmetry of a distribution)**, and kurtosis** (measures the thickness of the tail ends of a distribution).

Check out the graph and working for probability distribution for two dice here:



**NORMAL DISTRIBUTION EXPLAINED**

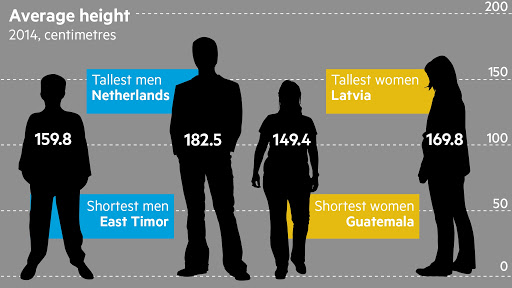
Out of the many probability distributions, the most widely used probability distribution is known as “**normal distribution**”. So, let us understand normal distribution now.

As mentioned before, if we plot the probability distribution and it forms a bell-shaped curve then the variable has normal distribution. Here, **the mean, mode and median of the sample are equal.**

This is an example of a normal distribution bell-shaped curve for two dice:



Some more examples whose variables are close to normally distribution are:

* Height of a population
* Blood pressure of adult human
* Position of a particle that experiences diffusion
* Measurement errors
* Residuals in regression
* Shoe size of a population
* Amount of time it takes for employees to reach home
* A large number of educational measures

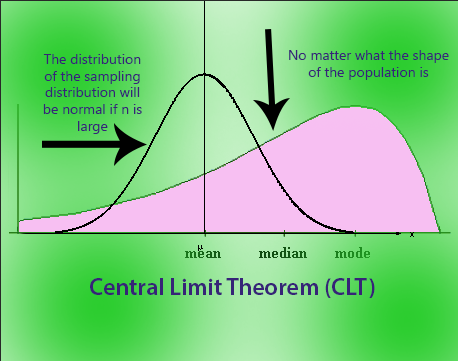
Moreover, there are many variables around us which are normal with a **x% confidence; x < 100.**

As mentioned before a normal distribution is a distribution that is solely dependent on two parameters of the data set: **mean and the standard deviation** of the sample, both of which have been explained above.

From a large variety of distributions, normal distribution is one of the **simplest** which makes it extremely easy for statisticians and hence any variable that exhibits normal distribution is feasible to be **forecasted with higher accuracy**.

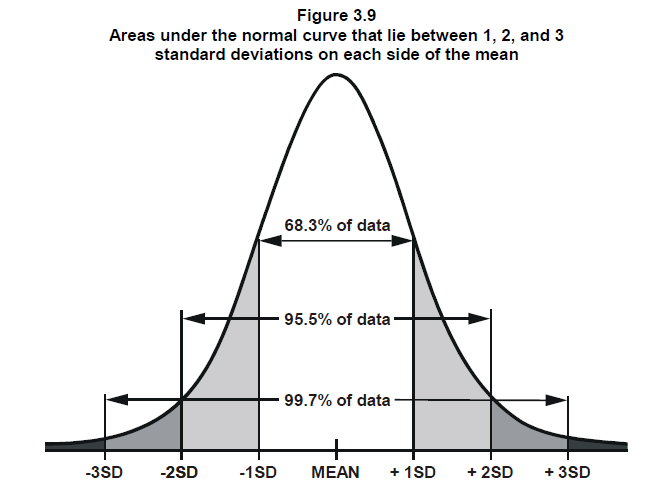
Now, an amazing feature of these distributions is that once you find the probability distributions of most of the variables in nature then they all approximately follow a normal distribution. Normal distribution is simple to explain because the mean, mode, and median of the distribution are equal and the **skewness is zero**. Moreover, we only require the mean and standard deviation to explain the entire distribution.

**What is the logical explanation for normal distribution?**

Let us consider that there is a random variable, such as the height of a sample of human population has a mean **m** and standard deviation **s**. Then, we would start gathering samples to represent the random variable let’s say ‘H’ for height, where each sample would have its own mean. Now, following the same steps as probability distribution, we would start repeating the experiment for more samples and start calculating the mean of each sample. After doing this, we will notice that the samples mean would have its own probability distribution as the number of samples are increased. This distribution will ultimately end up being very close to normality. Take the dice for example, here we see a triangle like shape which can be considered as normal distribution just with two dice but as we increase the samples to lets say 5 dice, it will get closer to normality and we will be able to see a more distinct normal distribution curve. This is known as the [**Central Limit Theorem**](https://www.investopedia.com/terms/c/central_limit_theorem.asp) (extremely important in normal distribution!).

Coming to the **main characteristics** of a typical normal distribution density function:

* **Mean** is the centre of the curve. This is the **highest point** of the curve as most of the points are at the mean.
* There is an **equal number of points** on each side of the curve (because of which skewness is zero).
* The **total area** under the curve is the total probability of all the values that the variable can take which is therefore **100%**.



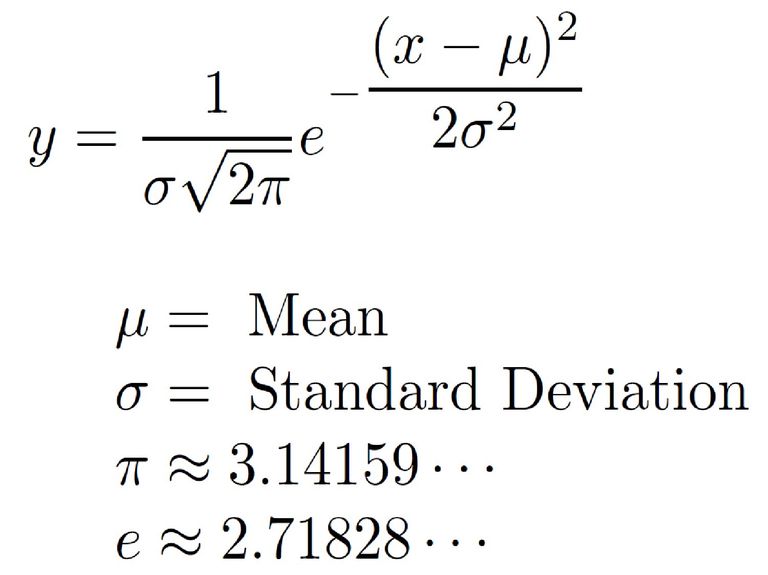
* Approximately **68.2%** of all of the points are within the range -1 to 1 standard deviation.
* About **95.5%** of all of the points are within the range -2 to 2 standard deviations.
* About **99.7%** of all of the points are within the range -3 to 3 standard deviations.

This allows us to easily estimate how **volatile a variable** is and to some level the prediction of how volatile it would be.

For example, there is a 68.3% chance that the value of the variable will be within the white range which on further giving it values can become more accurate.

**Normal Probability Distribution Function**

The probability density function of the normal distribution is:



The probability density function is the **probability of continuous random variable** taking a value. As we know, normal distribution is a bell-shaped curve where mean=mode=median.

So, if you plot the probability distribution curve using the **calculated** probability density function then the area under the curve for a given range gives the probability of the target variable being in that range.

This probability distribution curve is based on a probability distribution function which itself is computed on many parameters such as **mean, or standard deviation** of the variable.

We could use this probability distribution function to find the relative chance of a random variable taking a value within a range. For example, we could record the daily returns of a stock for lets say two years and group them into their appropriate categories and then find the probability of the stock making 20–40% gain in the future. Here is an example of the same:



Here, we calculated the return percentage of the stock by dividing the current month's price by the prior month's price and subtracting the number 1 this result before converting it into percentage. We, then calculated its mean and standard deviation and further on used the formula for normal distribution on excel and then plotted its graph. We see that it forms a normal distribution curve from which we can predict the gain %s.

**Problems:** The normal distribution is overused in a lot of predictive projects, mainly because it is simple and well-understood but it also has some flaws like we **cannot assume** that the stock price follows normal distribution as the price cannot be negative (we follow log of stock price so that its value is never below zero).

One way to make our sample distribution is **Linear Transformation** which focuses on computing **Z-scores.** The formula and working for the same is shown in our stock returns example.

However, the **daily returns can be negative** as seen in our example thus, the returns can at times follow a normal distribution. So, one cannot always assume that the data we have will follow normal distribution without further analysing it. Thus, the most important point to note is to understand and identify your data before you make the curve.

**Sources:**

* <https://statisticsbyjim.com/>
* <https://www.investopedia.com/terms/n/normaldistribution.asp#:~:text=Normal%20distribution%2C%20also%20known%20as,appear%20as%20a%20bell%20curve.>
* <https://www.youtube.com/watch?v=KjRj0Q8SIMI>
* <https://medium.com/fintechexplained/ever-wondered-why-normal-distribution-is-so-important-110a482abee3>
* <https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/more-on-normal-distributions/v/introduction-to-the-normal-distribution>