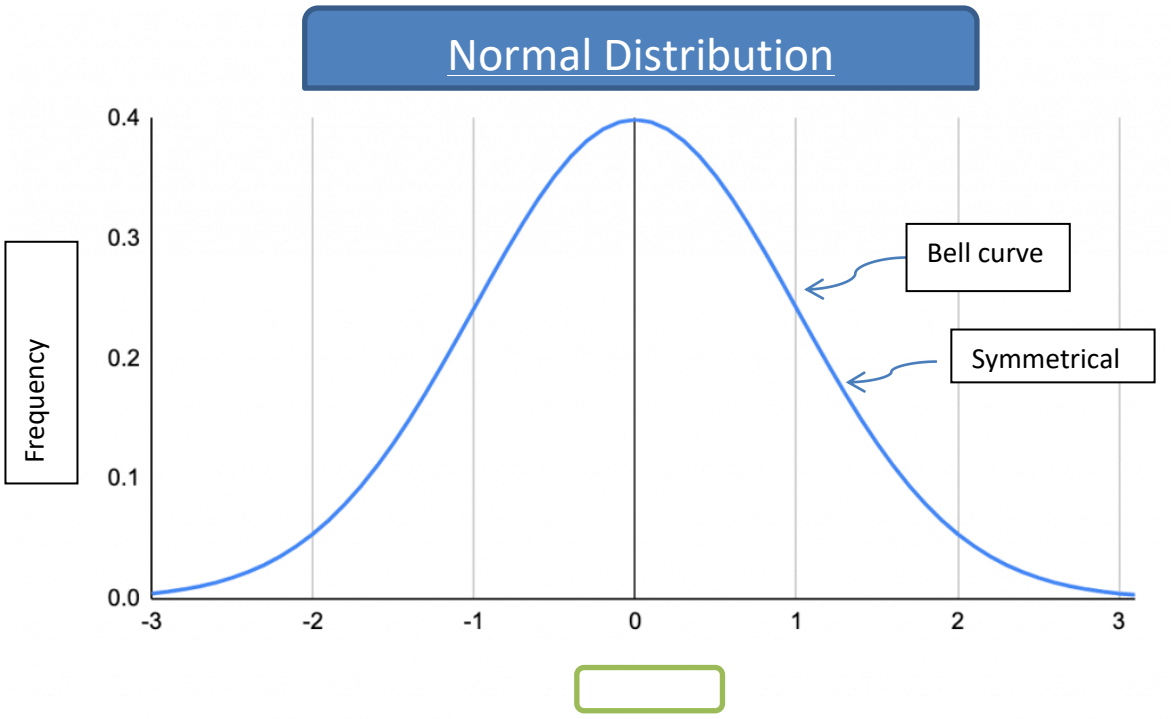


- **Normal Distribution** :

  Normal Distribution or Gaussian Distribution is a probability distribution that is symmetric about the mean. It shows the data nearer to the mean has more frequency than data far from mean.

  In graph when we plot normal distribution it looks similar to bell curve.

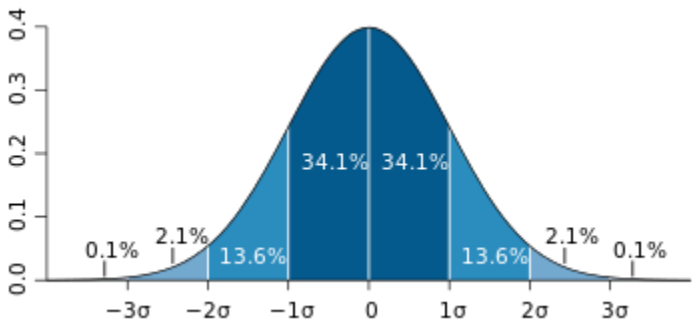
- Normal Distribution gives a lot of assumption of data. It is important because it fits a lot of natural phenomena like measurement of error, blood pressure, IQ score and it has a lot of mathematical properties.

  Most of the data set follows normal distribution.

  Eg – Age, Weight, Height, IRIS Date

  Bell curve : When we smoothen a histogram with kernel density estimation we get a bell curve.

- Empirical Rule of Normal / Gaussian Distribution :



Empirical rule also known as 3σ or 68-68-99.7 rule.

- Empirical rule state that 68% of the observation falls within first standard deviation (μ ± σ), 95% of observation falls within first two standard deviations (μ ± 2σ) and 99.7% of the observation falls within first three standard deviations (μ ± 3σ).

Standard Normal Distribution :

- The Standard Normal Distribution (z distribution) is a normal distribution where mean is 0 and standard deviation is 1.

Any normal distribution can be standardized by converting it's values into z-score. Z-score tells you that how many standard deviation the values are away from mean.

- Z-Score :

The formula to calculate Z-score : **$z = (x - \mu) / (\sigma / \sqrt{n})$**

As we are applying Z-score to each and every value, we are considering n = 1.

So the derived formula is : $z = (x - \mu) / \sigma$

- Eg.

$\chi = \{1,2,3,4,5\}$, $\mu = 3$, $\sigma = 1$

If we will apply z-score to this data set then,

$\gamma = \{-1.414, -0.707, 0, 0.707, 1.414\}$

Why we apply Z-Score? ['Standardization']

- If the data points will be of different unit then Machine Learning algorithm will take more to process them.

So we apply Z-score on these data set to bring the features to the same scale, for which the calculation will become easy and we can quickly get result.

- This entire process is called '**Standardization'**. This process allows us to compare scores between different types of variables.

X => Normal Distribution (μ , σ)

⇩ Z-score

Y => Standard Normal Distribution (μ = 0 , σ = 1)

- **Feature Scaling** :

It is a step of Data pre-processing that is applied to independent variables or features of data. It basically helps to normalize the data within particular range.

Feature scaling techniques are **Standardization** and **Normalization**.

- In **Normalization** we shift the mean to 0. The purpose of Normalization to scale down the values in between 0 and 1.

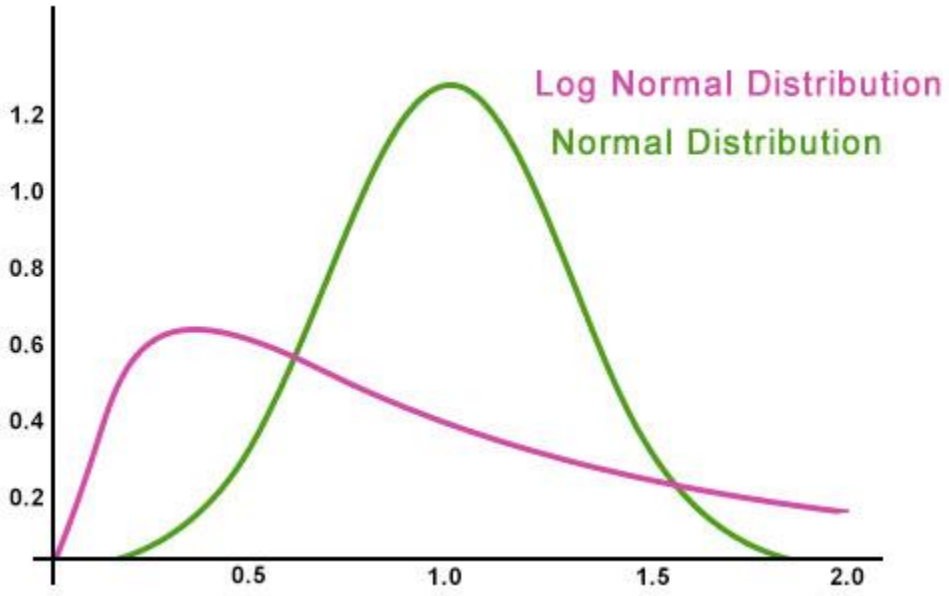
- MinMax Scaler :

  MinMax Scaler is a Normalization technique with the help of it we can convert the data which will range in between 0 to 1.

  In MinMax Scaler the distribution will get changed. So in some of the cases it may get failed.

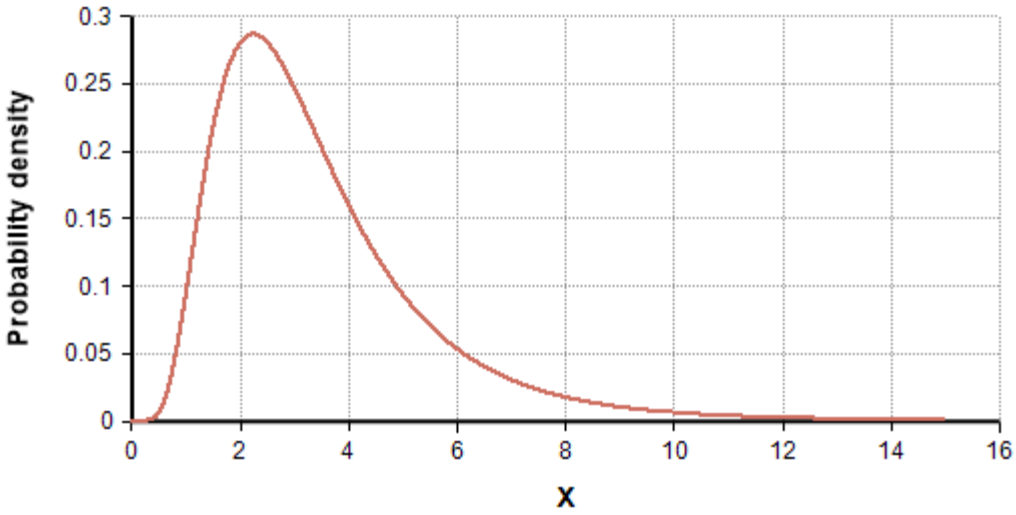
- **Log Normal Distribution** :

  A log-normal distribution is a continuous distribution of random variable whose natural logarithm is normally distributed.



  Eg – Wealth distribution, Length of comments,

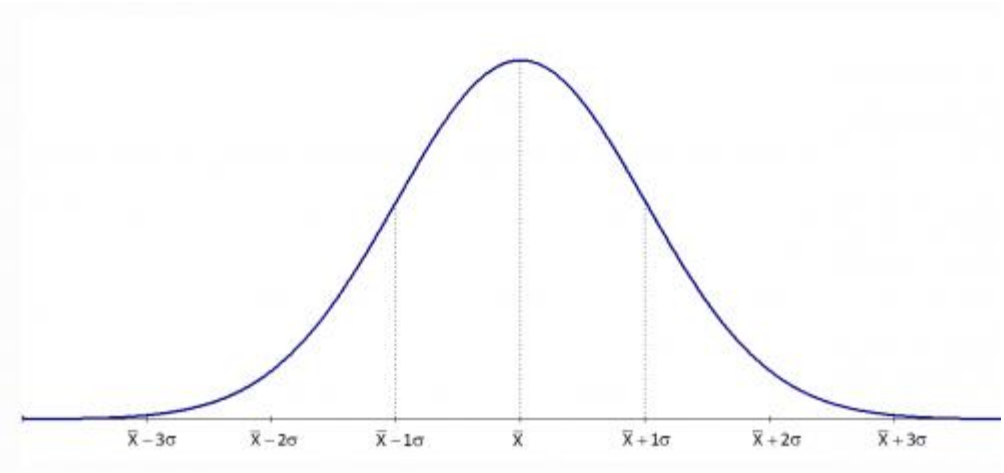
  It is a distribution which is skewed to right. So in the higher scale there are outliers.



$X \approx$ Log Normal Distribution

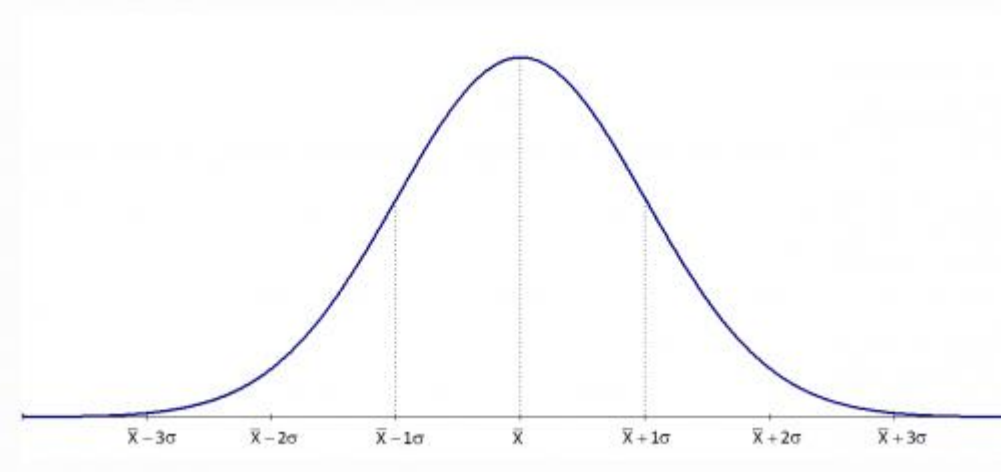
If we apply natural log over X then it will be log normally distributed.

$Y = \ln(X)$

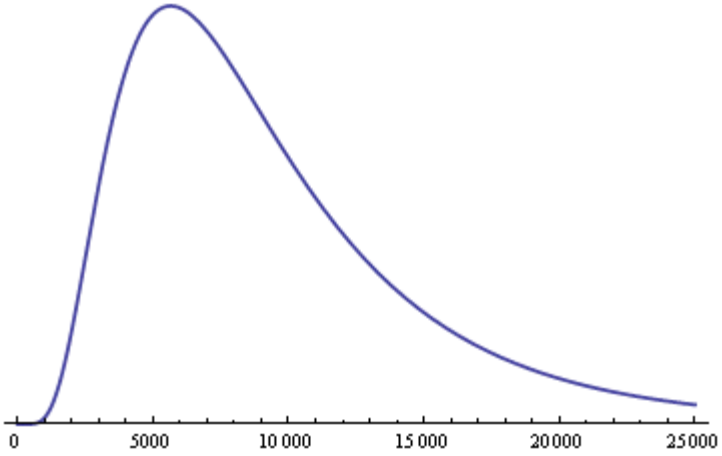


If it is normal distribution we can say X is log normally distributed.

- So the Inverse is also true.



X = exp(Y)



Examples of log normal distribution – Scores of Bat's man

- To check if X ≈ Log Normally Distributed, we need to ensure we are getting Normal distribution out of Y = ln(X).

- X = {1,2,3,4,5,6,7} μ = 4, σ = 1

What is the percentage of score that falls above 4.25?

Z-score = $\frac{x-\mu}{\sigma} = \frac{4.25-4}{1}$ = 0.25 ≈ 0.59871 ≈ 59%

⇨ 1 - .59 = .41 ≈ 41%

Z-score is helpful to find out how much away the point is away from mean.

- X = {1,2,3,4,5,6,7} μ = 4, σ = 1

  What is the percentage of score that falls below 3.75?

  Z-score = $\frac{x-\mu}{\sigma}$ = $\frac{3.75-4}{1}$ = -0.25 ≈ 0.40129 ≈ 40%

- What is the percentage of score that falls between 4.75 & 5.75 ?

  μ = 4, σ = 1

  Z-score = $\frac{x-\mu}{\sigma}$

  For 5.75 = $\frac{5.75-4}{1}$ = 1.75 ≈ 0.95994

  For 4.75 = $\frac{4.75-4}{1}$ = 0.75 ≈ 0.77337

  Result = 0.95994 – 0.77337 = 0.18657 ≈ 18%

- In India average IQ is 100 with a standard deviation of 15. What is the percentage of population would you expect to have an IQ

  1. Lower than 85
  2. Higher than 85
  3. Between 85 and 100

  μ = 100, σ = 15

  1. Lower than 85

  Z-score = $\frac{x-\mu}{\sigma}$ = $\frac{85-100}{15}$ = 1 ≈ .84134 ≈ 84%

  2. Higher than 85

  Z-score = $\frac{x-\mu}{\sigma}$ = $\frac{85-100}{15}$ = 1 ≈ .84134 ≈ .84 ≈ 84%

  ⇨ 1-..84 = .16 ≈ 16%

  3. For 100 : $\frac{100-100}{15}$ = 0 ≈ 0.5

  For 85 : $\frac{85-100}{15}$ = $\frac{-15}{15}$ = -1 ≈ .15866

  Between 85 and 100 : 0.5 – 0.15866 = .34134 ≈ 34%