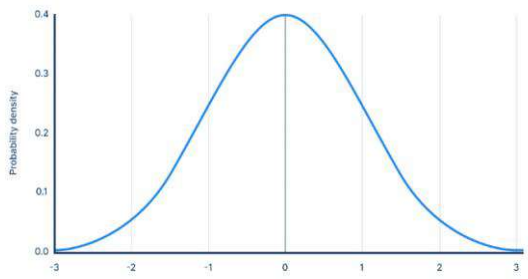
NORMAL DISTRIBUTION

Normal distribution, also known as **Gaussian distribution**, is a probability distribution that is commonly used in statistical analysis. It is a continuous probability distribution that is symmetrical around the mean, with a **bell-shaped** curve.



NOTE-The normal distribution is characterized by two parameters: the **mean (μ)** and the **standard deviation (σ)**. The mean represents the **centre of the distribution**, while the standard deviation represents the **spread of the distribution**.

Denoted as: X~N(μ, σ)

$\boldsymbol{\mu = mean}$ $\boldsymbol{\sigma = standard\ devation}$

- Tail
- **Asymptotic** in nature
- Lots of points near the mean and very few far away

Why is it so important?

Commonality in Nature: Many natural phenomena follow a normal distribution, such as the heights of people, the weights of objects, the IQ scores of a population, and many more. Thus, the normal distribution provides a convenient way to model and analyse such data.

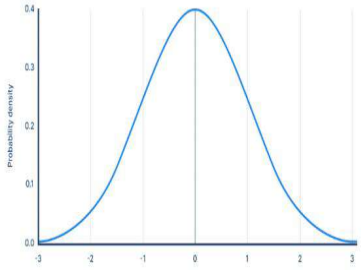
PDF Equation of **Normal Distribution** $y=F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma^2})^2}$

Impact of mean and standard devation-> https://samp-suman-normal-dist-visualize-app-lkntug.streamlit.app

Standard Normal Variate/ Standard Normal Distribution(Z)

What is Standard Normal Variate?

A Standard Normal Variate(Z) is a standardized form of the normal distribution with mean = 0 and standard deviation = 1.

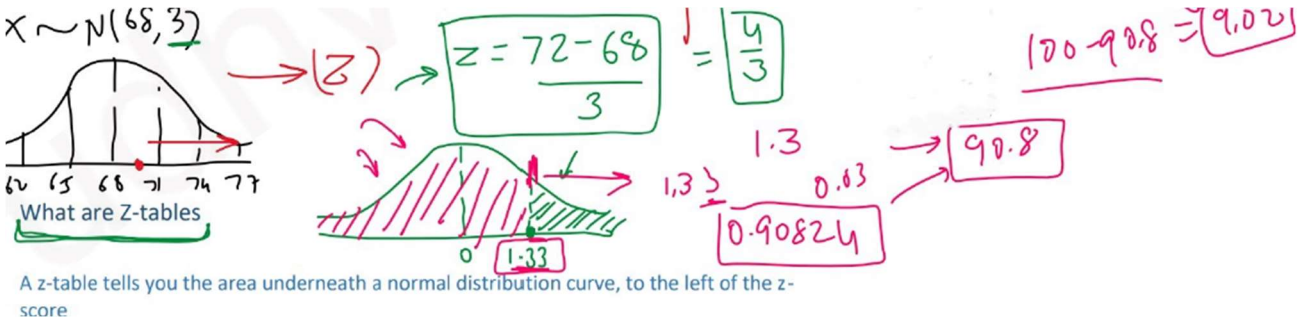


Standardizing a normal distribution allows us to compare different distributions with each other, and to calculate probabilities using standardized tables or software. **Equation:** $y=\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

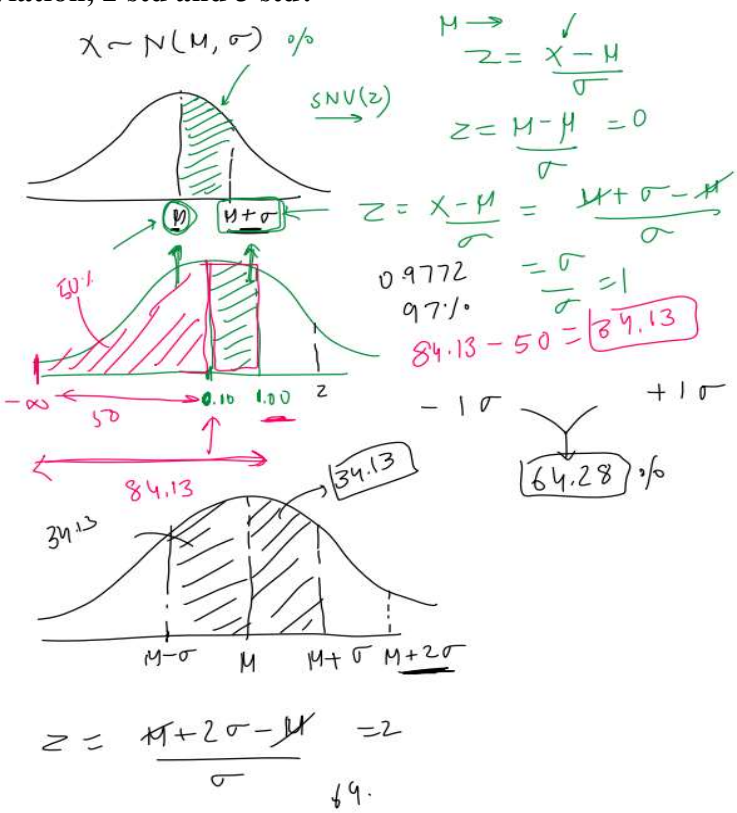
$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1(x)^2}{2}}$ where **mean=0 and standard devation=1**

HOW TO TRANSFORM A NORMAL DISTRIBUTION TO STANDARD NORMAL VARIATE

1. Suppose the heights of adult males in a certain population follow a normal distribution with a mean of 68 inches and a standard deviation of 3 inches. What is the probability that a randomly selected adult male from this population is taller than 72 inches?



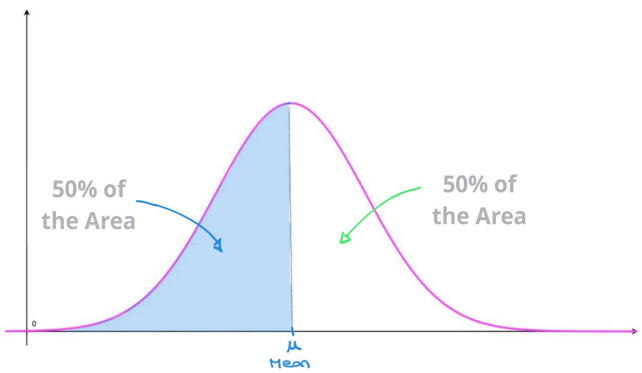
Q2-For a Normal Distribution X~(u,std) what percent of population lie between mean and 1 standard deviation, 2 std and 3 std?



PROPERTIES OF NORMAL DISTRIBUTION

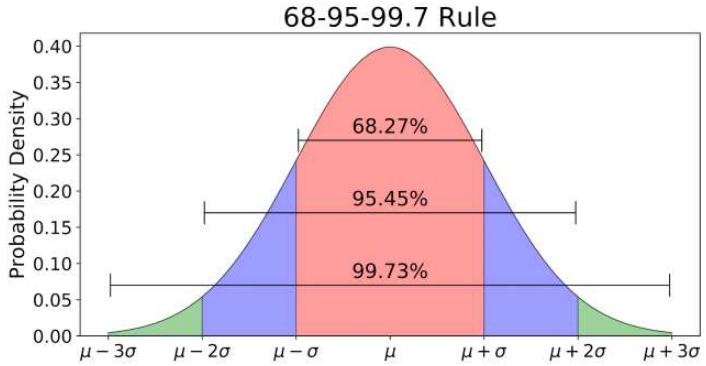
1.symmetricity

The normal distribution is **symmetric about its mean**, which means that the probability of observing a value above the mean is the same as the probability of observing a value below the mean. The bell-shaped curve of the normal distribution reflects this symmetry.



2.measure of central tendencies are equal: mean=median=mode

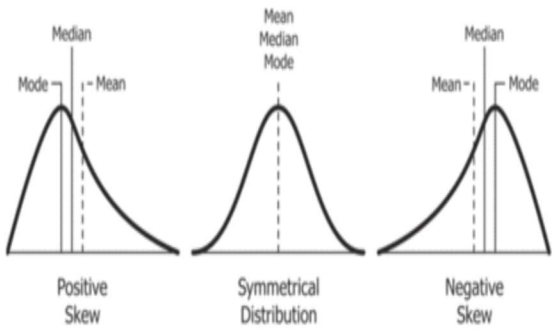
3.Empirical Rule: The normal distribution has a well-known empirical rule, also called the **68-95-99.7 rule**, which states that approximately 68% of the data falls within one standard deviation of the mean, about 95% of the data falls within two standard deviations of the mean, and about 99.7% of the data falls within three standard deviations of the mean.



4. The area under the curve = 1

Skewness

- A normal distribution is a bell-shaped, symmetrical distribution with a specific mathematical formula that describes how the data is spread out**. Skewness indicates that the data is not symmetrical, which means it is not normally distributed.**
- Skewness is a measure of the asymmetry of a probability distribution. It is a statistical measure that describes the degree to which a dataset deviates from the normal distribution.
- In a symmetrical distribution, the mean, median, and mode are all equal. In contrast, in a skewed distribution, **the mean, median, and mode are not equal**, and the distribution tends to have a longer tail on one side than the other.(mode<median<mean)
- Skewness can be positive, negative, or zero. A positive skewness means that the tail of the distribution is longer on the right side, while a negative skewness means that the tail is longer on the left side. A zero skewness indicates a perfectly symmetrical distribution.



Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).

The greater the skew the greater the distance between mode, median and mean.

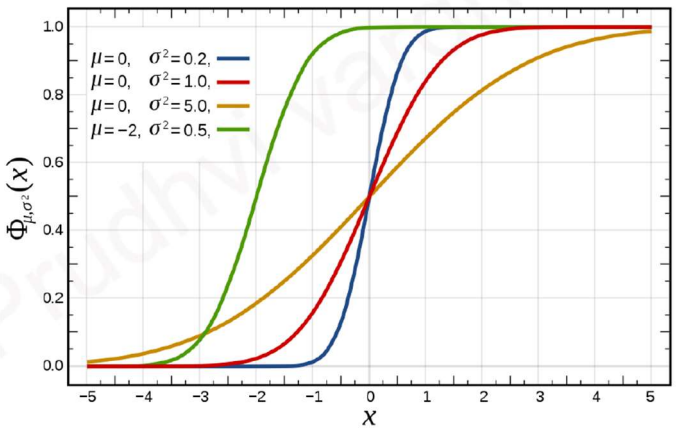
$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{(x-\bar{x})}{s}\right)^3$$

0 -> perfect normal distribution

-0.5to 0.5 ->All most symmetric normal distribution

1 to -1 ->highly symmetrical

CDF of Normal Distribution



$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Normalisation vs. Standardisation

- **Normalisation** is suitable to use when the **data does not follow Gaussian Distribution principles**. It can be used in algorithms that do not assume data distribution, such as K-Nearest Neighbors and Neural Networks.
- On the other hand, **standardisation** is beneficial in cases where the **dataset follows the Gaussian distribution**. Unlike Normalization, Standardisation is not affected by the outliers in the dataset as it does not have any bounding range.
- Applying Normalization or Standardisation depends on the problem and the machine learning algorithm. There are no definite rules as to when to use Normalization or Standardisation. One can fit the normalized or standardized dataset into the model and compare the two.
- It is always advisable to first fit the scaler on the training data and then transform the testing data. This would prohibit data leakage during the model testing process, and the scaling of target values is generally not required.

Normalisation	Standardisation
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers
It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed
It is also known as Scaling Normalization	It is also known as Z-Score

USE IN DATA SCIENCE

1. Outlier detection
2. Hypothesis Testing
3. Central Limit Theorem
4. Assumptions on data for ML algorithms -> Linear Regression and GMM