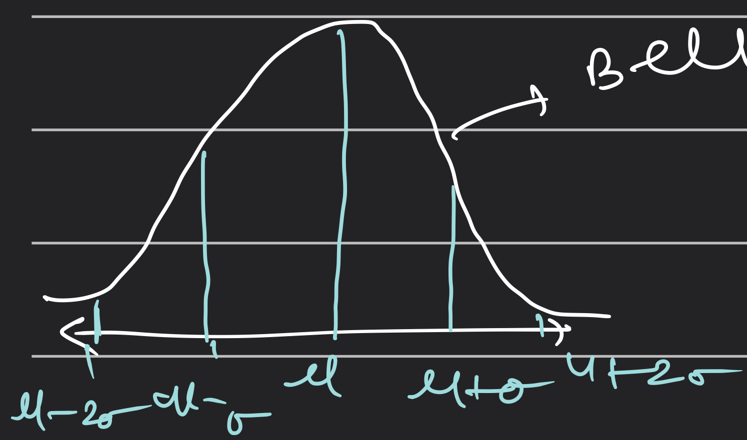
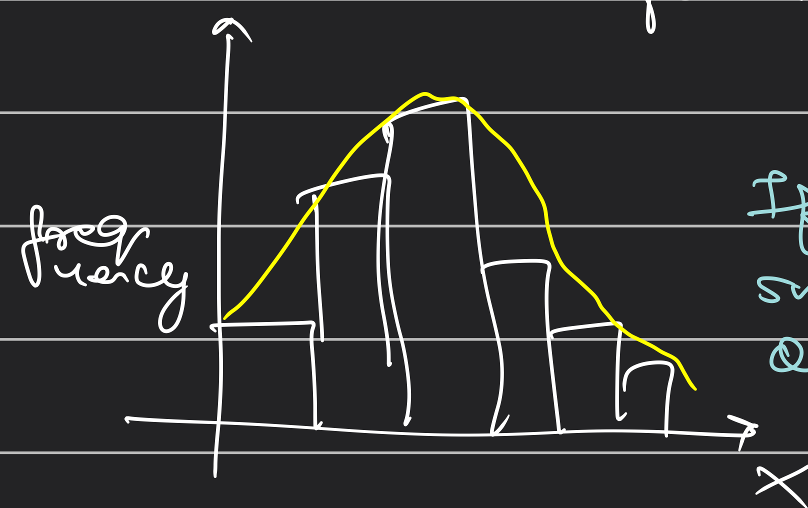
Normal / Gaussian Distribution



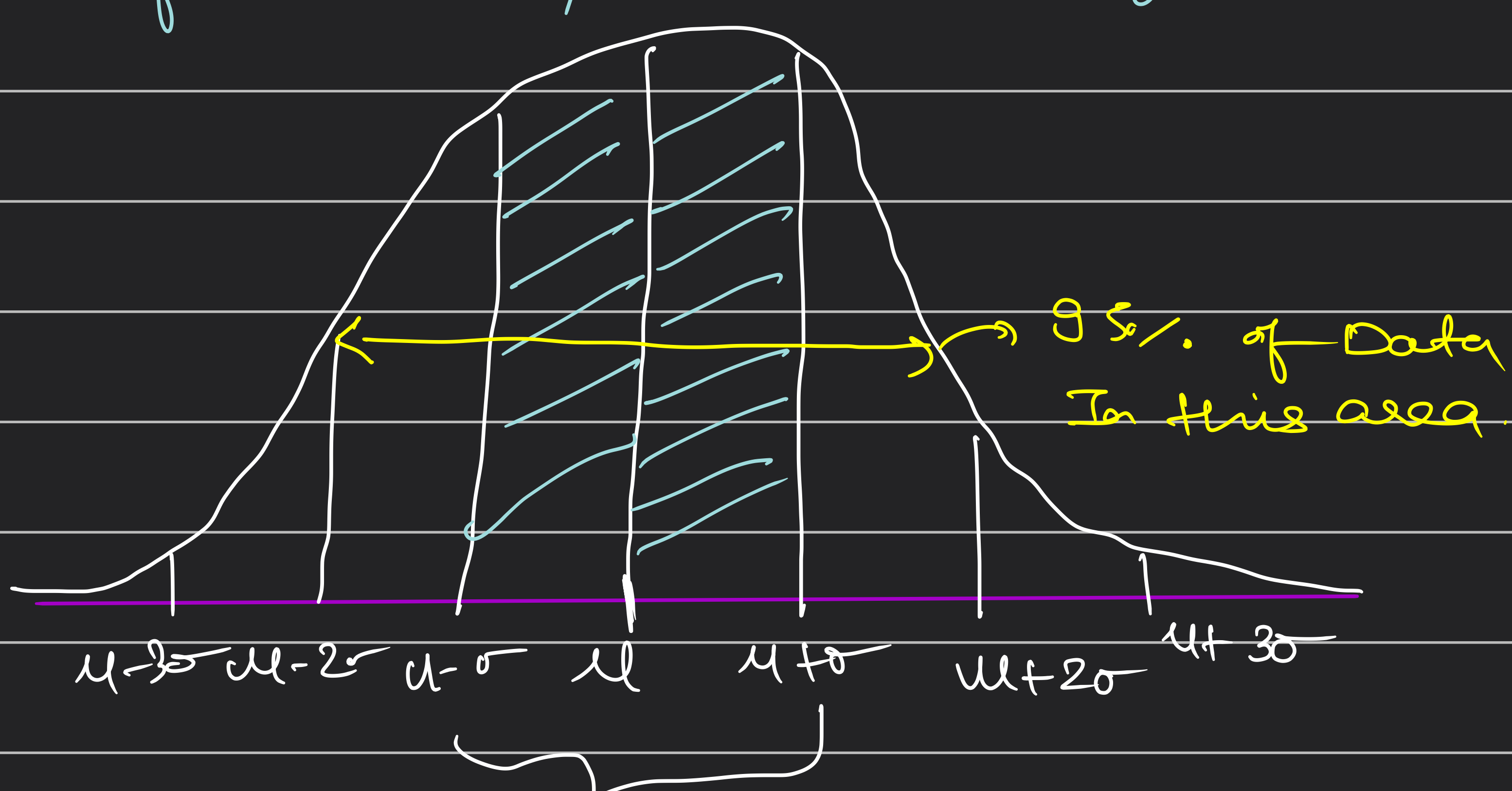
Bell Curve → How do we get it

⇓ We first make a histogram



If we do smoothening of our histogram through a technique called Kernel Density Estimator this smoothning results in Probability Density function which is our Bell curve.

Left Area of curve & Right area of curve are symmetrical. 50% of numbers on left side of mean & 50% to right side of mean



- Emperical Rule [68-95-99.7]

According to emperical rule 68% of the data will fall in 1SD, 95 in 2SD'S and 99.7 in 3SD'S. whenever the random variable follow the Gaussian or Normal Distribution.

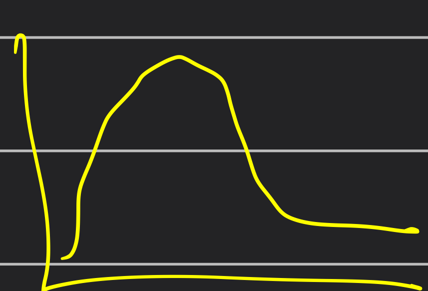
Central Limit theorem

So whenever you take random samples (multiple) from a population which could or could not be normally distributed, may be log distributed (skewed) towards right, now if you take mean of these samples then the distribution of all such means of the samples will approximate a normal distribution.
The sample size for each sample should be >= 30.

Distribution could be any type, sample size >=30, then mean of such samples when plotted will result in normal distribution.

$x = \{65, 72, 83 \ldots \ldots 94 \ldots\}$

$n >= 30 \rightarrow$ sample size



$\rightarrow \{x_1, x_2, x_3, \ldots\} \rightarrow \bar{x}_1$ mean

$\rightarrow \{x_1, x_3, x_7 \ldots\} \rightarrow \bar{x}_2$

$\rightarrow \{x_5, x_{11}, x_{19} \ldots\} \rightarrow \bar{x}_3$

Now If we plot these $\bar{x}_1, \bar{x}_2, \bar{x}_3$ we will get a Normal distribution / Gaussian curve.

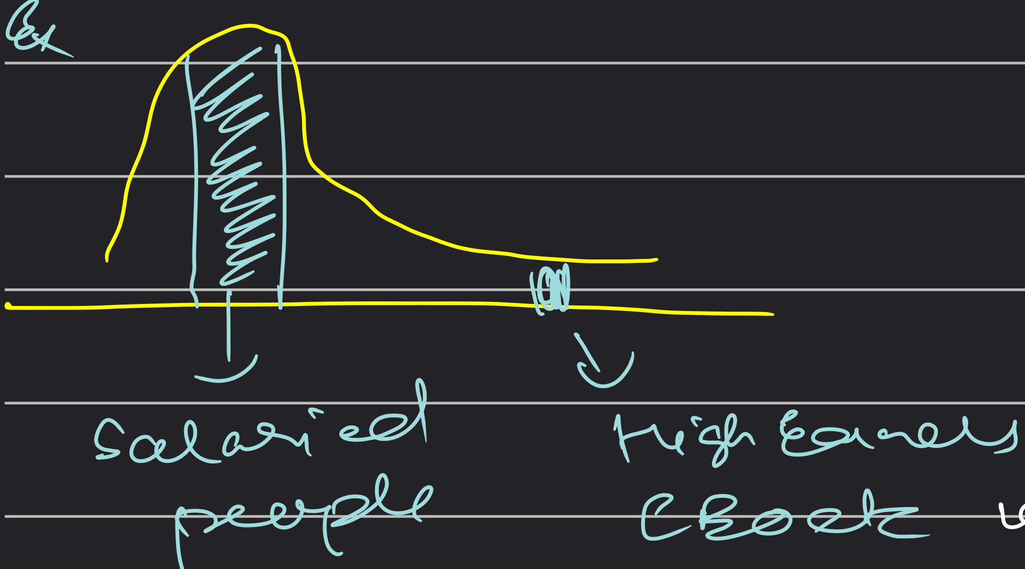
Log - Normal - distribution.

Looks like a gaussian distribution but is elongated towards right or we call it right skewed.

$\log_e$ = natural log

If $X \underset{\text{"belongs to"}}{\sim}$ log normal distribution, then $Y \sim \ln(X)$
then plot points in Y then I will get a normal distribution. We are applying this natural log to each and every element of X and then plotting it.

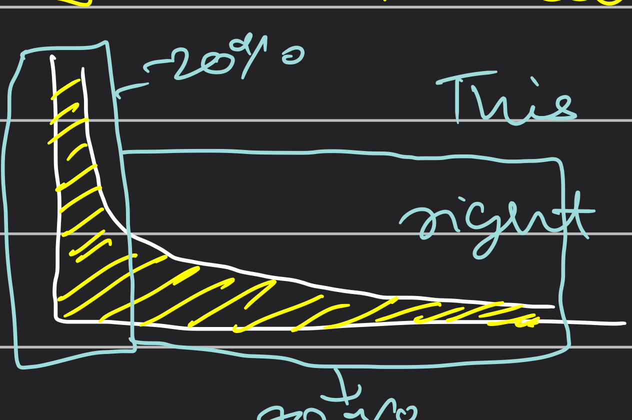
Ex



The log normal distribution is a probability distribution where logarithm of the data follows a normal distribution.
Used for variables that are naturally positive & skewed like income or stock prices.
Exponent of normal distribution will result in log normal distribution

How are these things utilized?
In ML $\boxed{\text{linear regression}}$ works wonders when data is normally distributed. Even if you have log-normal distributed data you can make it normally distributed by using log on it.

Power Law Distribution



This distribution has long tail towards right and left have few that dominates (also like 80-20 rule)

Eg① 20% of team are responsible for winning 80% of matches

Eg② 80% of entire wealth is owned by 20% of people.

Power law distribution is to describe phenomenon where few occurrences have a very high frequency. ex is distribution of wealth.

So probability of event is inversely proportional to its size.

one variable is the power of other variable.

Can we transform to Normal Distribution?

Ans → Box Cox transform.

- Pareto Distribution follows power law distribution.

Pareto Distribution. [80-20 % Rule]

⇓ Non Gaussian

- Z Score

Helps us find out how far away (S.D) is a value from mean.

$SD = \sigma = 1$

$$Z\,score = \frac{x_i - \mu}{\sigma}$$

$\mu = 4$

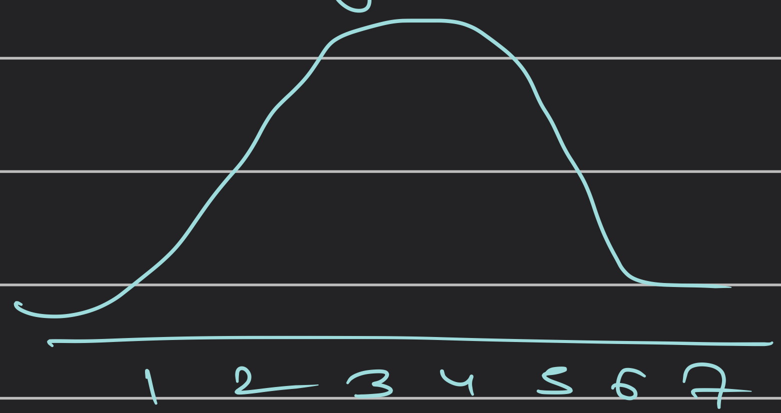
$x_i = 4.75$

$$Z\,score = \frac{4.75 - 4}{1} = 0.75\ SD \text{ towards right}$$

If -ve then towards left.

Standard Normal Distribution

lets say $\mu = 4$ $\sigma = 1$



Zscore on Each value
$\{-3, -2, -1, 0, 1, 2, 3\}$
Now this distribution has $\mu = 0$ & $SD = \sigma = 1$

Standard Normal Distribution is what we get when we apply Zscore on our Normal Distribution

A random variable X will belong to Standard Normal distribution if $\mu = 0$ & $\sigma = 1$ for that distribution.

Normalization

rescaling numeric data to standard range typically between 0 and 1

Is useful when features have different scales. It ensure all the features contribute equally to analysis and prevents one feature from dominating others due to large scale.