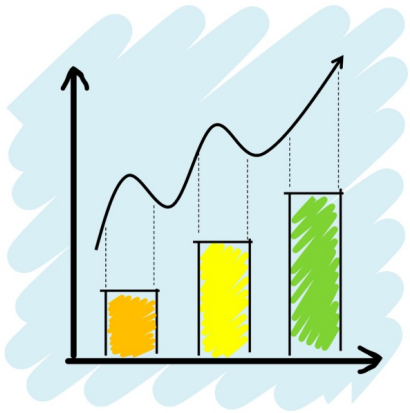


Statistics

ML-ops - workflow

- R
- python

1. define problem: — hypothesis
2. collect the data
3. organize the data
4. cleanse the data
5. perform operation / algorithm : find result
6. hypothesis testing
   - distributions - T, F, chi.. 2
   - success
   - failure

x	y
1	3
2	5
3	7
4	9
5	11

known data

$y = 2x + 1$

formula = model (lm) → R

predict() → R

unknown data

distribution

$x = 1$ to $x = 100$

$y = \{ \ldots \ldots \}$

python → Sci-kit learn → LinearRegression ⇓ model

distribution function

```
def generate_values(x):
    return 2x+1
```

Distribution : function

↳ going to generate listing

- The distribution of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur
- When a distribution of categorical data is organized, you see the number or percentage of individuals in each group → factors → levels → unique values
- When a distribution of numerical data is organized
  - they're often ordered from smallest to largest, broken into reasonably sized groups (if appropriate)
  - then put into graphs and charts to examine the shape, center, and amount of variability in the data
- Types
  - Discrete probability distribution
  - Continuous probability distribution

Terminologies

- **Frequency distribution**
  - A frequency distribution is a table that displays the frequency of various outcomes **in a sample**.
- **Relative frequency distribution**:
  - A frequency distribution where each value has been divided (normalized) by a number of outcomes in a sample i.e. sample size.
- **Probability distribution**:
  - alias for *Relative frequency distribution*
  - indicates the way the total probability of 1 is distributed over **all** various possible outcomes

$x_1 = \{10, 20, 10, 15, 18, 20, 18, 10, 15\}$

frequency distribution

x	frequency
10	3
20	2
15	2
18	2
	9

Relative frequency distribution or probability distribution

x	frequency	relative frequency / probability
10	3	$p(10) = 3/9$
20	2	$p(20) = 2/9$
15	2	$p(15) = 2/9$
18	2	$p(18) = 2/9$
	9	$= \frac{3+2+2+2}{9} \rightarrow (1)$

9 → total outcomes

addition of all probabilities = (1)

$x = \{ 5, 8, 9, 7, 3, 4, 8, 9, 10, 2, 3, 5, 5 \}$

Discrete Probability Distribution

Discrete Probability Distribution (function)

- Is also called as probability mass function (PMF)
- The sum of all the individual probabilities must equal to 1
- Types
  1. Uniform distribution : same probabilities for all outcomes
  2. Binomial distribution : only two possible outcomes (n bernoulli trials)
  3. Negative Binomial distribution
  4. Poisson distribution
  5. Geometric distribution

① Binomial : probability of x success in n trials

② Negative Binomial : find no of trials for constant no of success

③ Geometric : find no of trials needed for first success

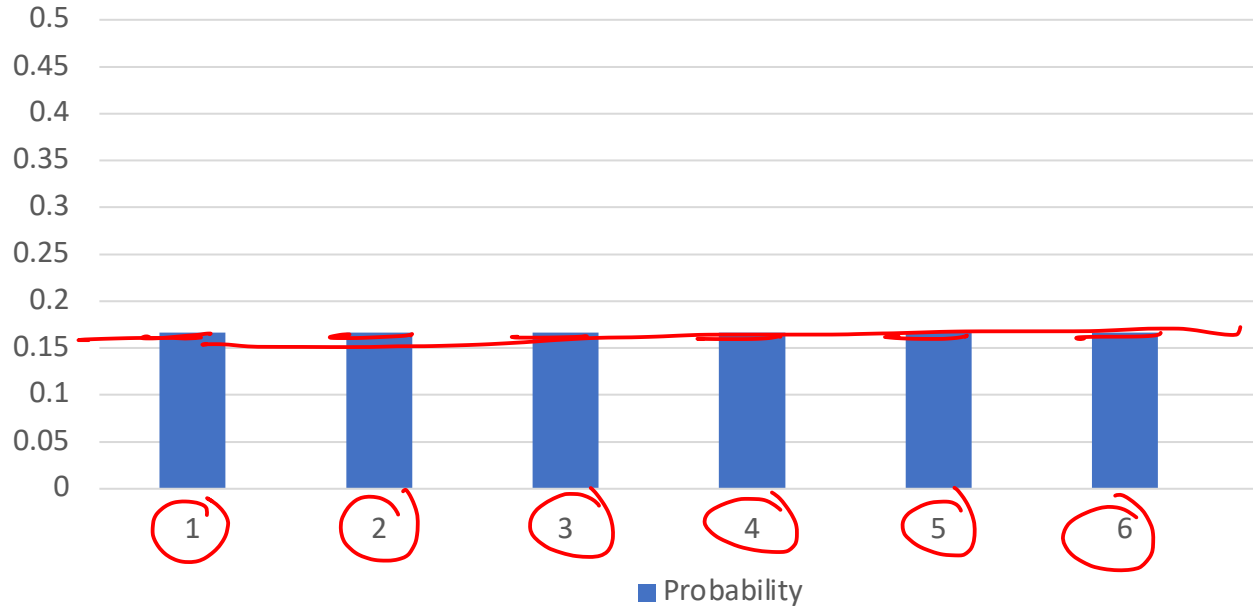
④ Poisson : find no of success in n trials

Uniform distribution

- The probabilities of each outcome are evenly distributed across the sample space
- E.g. Rolling a fair die has 6 discrete equally probable outcomes

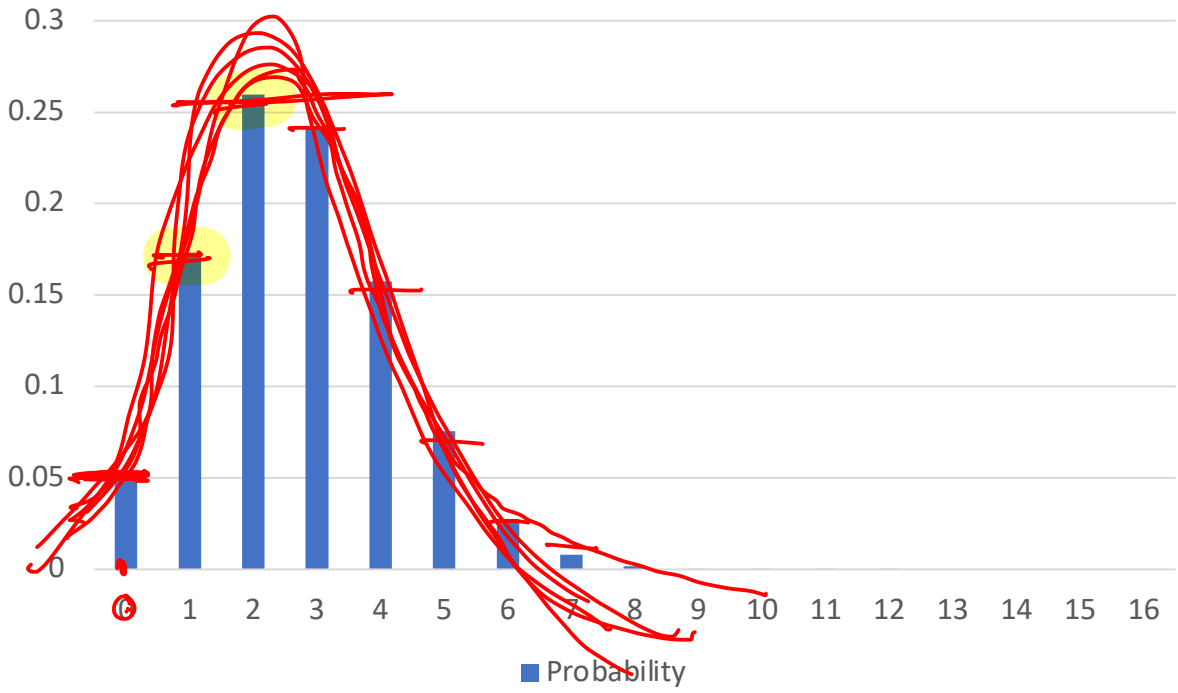
P(1) = 1/6
P(2) = 1/6
⋮
P(6) = 1/6

6/6 = 1



Binomial distribution

- Binomial means two discrete outcomes
- The outcomes of a trial are mutually exclusive
- Bernoulli Trial
  - When a trail is conducted there will be only two possible outcomes
  - Success or Failure
- A series of n trials follow binary distribution when
  - The probability of success (p) is a constant
  - Trials are independent of one another
- E.g. Flipping a coin



Binomial distribution

function

- Probability mass function
  - Generates probability of observing x successes in n trials
  - Probability of single trial is constant (p)

2 - 10 trials

1 = 10 = 9 = 1

2 = 10 = 0 = 0

3 = 10 ⇒ 10 = 0

1/3

1000

success

→ head

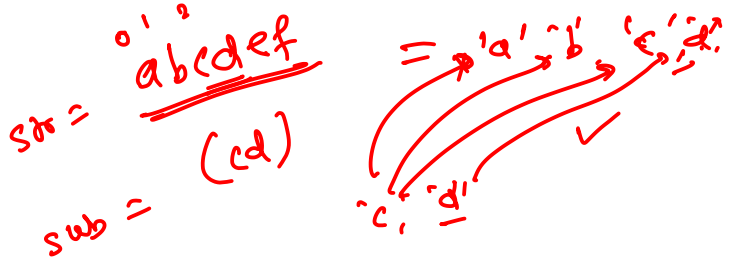
$$p(x:n,p) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

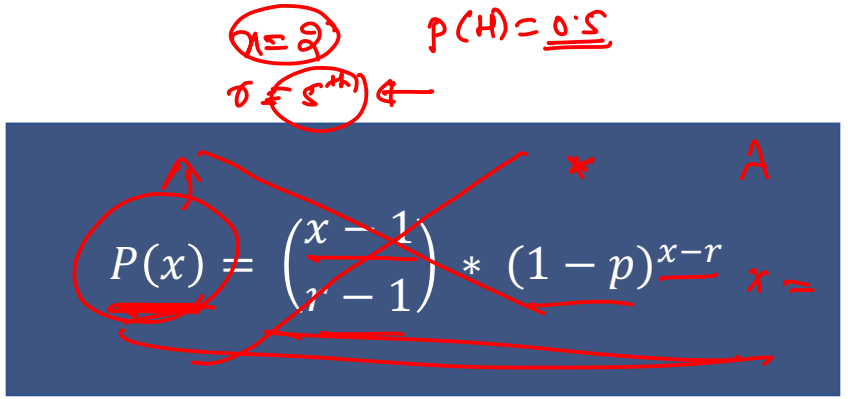
^nC_x $\binom{n}{x}$

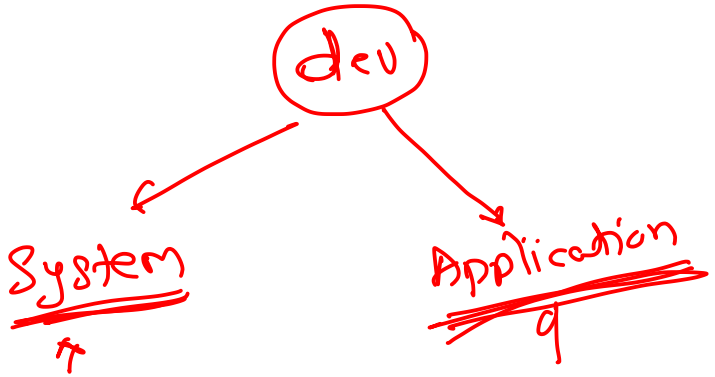
Negative Binomial distribution

- Negative Binomial Distribution is of number of trials needed to get a fixed number of successes
- Each trial results in one of the two possible outcomes
  - Success
  - Failure
- Probability of success = P(Success) = p (constant)
- Probability of failure = P(Failure) = 1 - p
- X represents the trial number of the rth success
- Probability Mass Function

$$P(x) = \binom{x-1}{r-1} * (1-p)^{x-r}$$







Geometric distribution

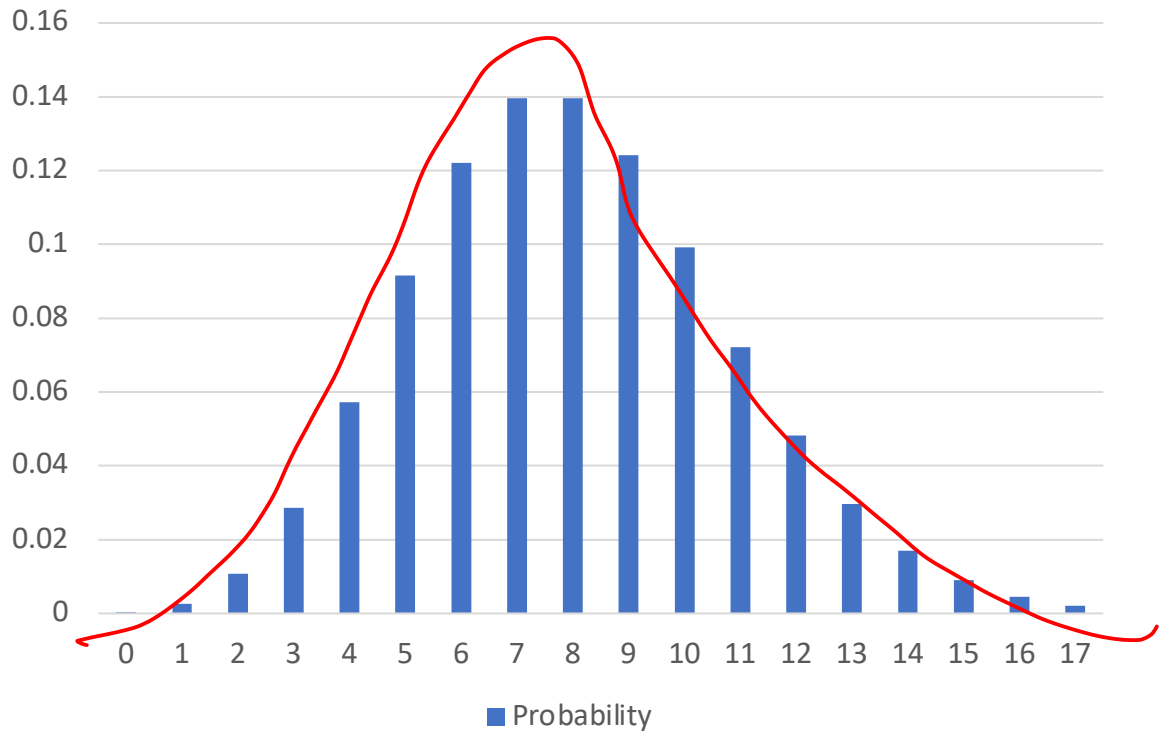
- Is a distribution of the number of trials needed to get the first success in the repeated Bernoulli trials
- Every trial results in one of two possible outcomes
  - Success
  - Failure
- P(Success) = p (stays constant)
- P(Failure) = 1 – p
- X represents the number of trails needed to get the first success
- For the first success to occur on the xth trial:
  - The first x – 1 trials must be failures
  - The xth trial must be a success
- Probability mass function

$$P(x) = (1-p)^{x-1} * p$$

Poisson distribution

- A Binomial distribution considers number of successes out of n trial
- A Poisson distribution considers number of successes per unit of time over the course of many units
- Terminology
  - X: Variable
  - E: Expectation (Mean) => E(X)
  - μ: Expectation (Mean) = E(X)
  - $\lambda = \frac{\#\ occurrences}{interval} = \mu$
- Probability Mass Function

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$



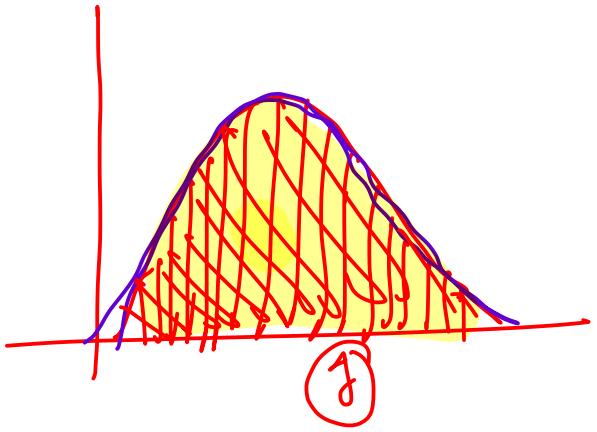
interval

Continuous Probability Distribution

Continuous Probability Distribution (PDF)

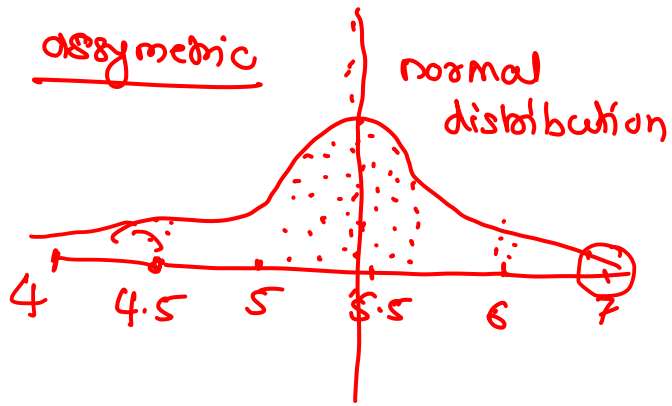
- Also called as probability density functions (PDF)
- Where the variable values are infinite → interval
- The area under the probability curve equals 1 →
  - Types
    - Normal distribution
    - Student T distribution
    - Chi-Squared distribution
    - Exponential distribution
    - Logistic distribution

↳ classification model evaluation (AUC)

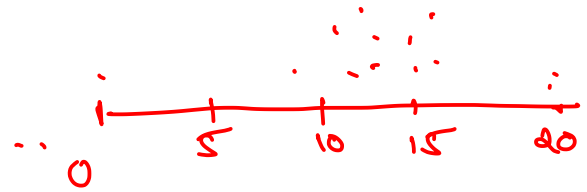


Normal Distribution

- Data sources tend to be around central value with no bias left or right
- Many real life data points follow normal distribution
- E.g.
  - Peoples height and weights
  - Population blood pressure
  - Test scores
  - Measurement errors
- Probability distribution function

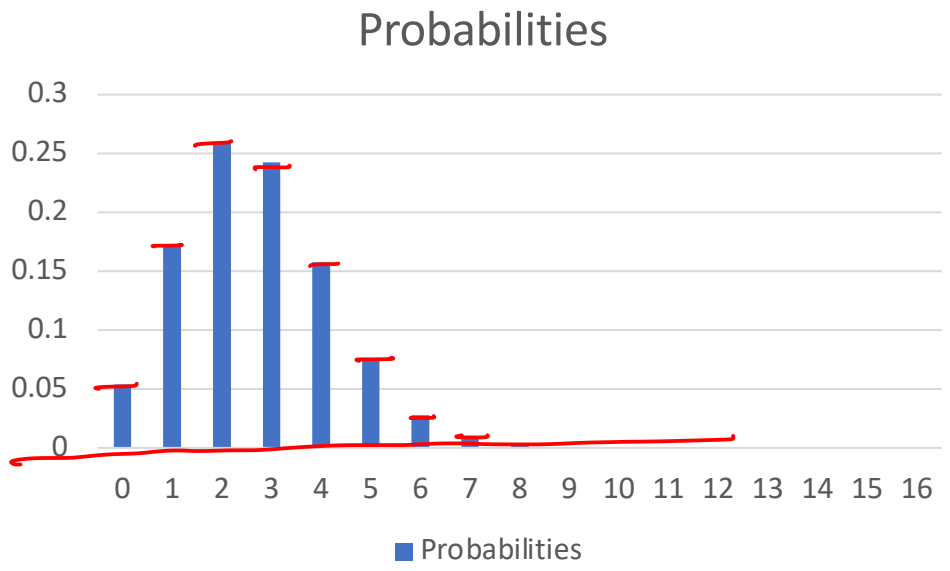


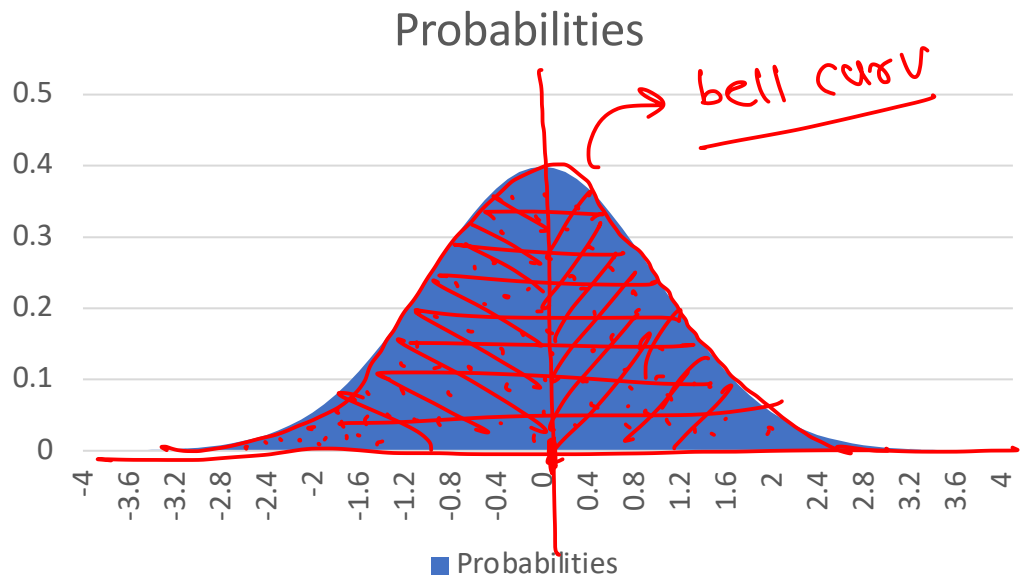
$$f(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}} \right)$$



Normal Distribution

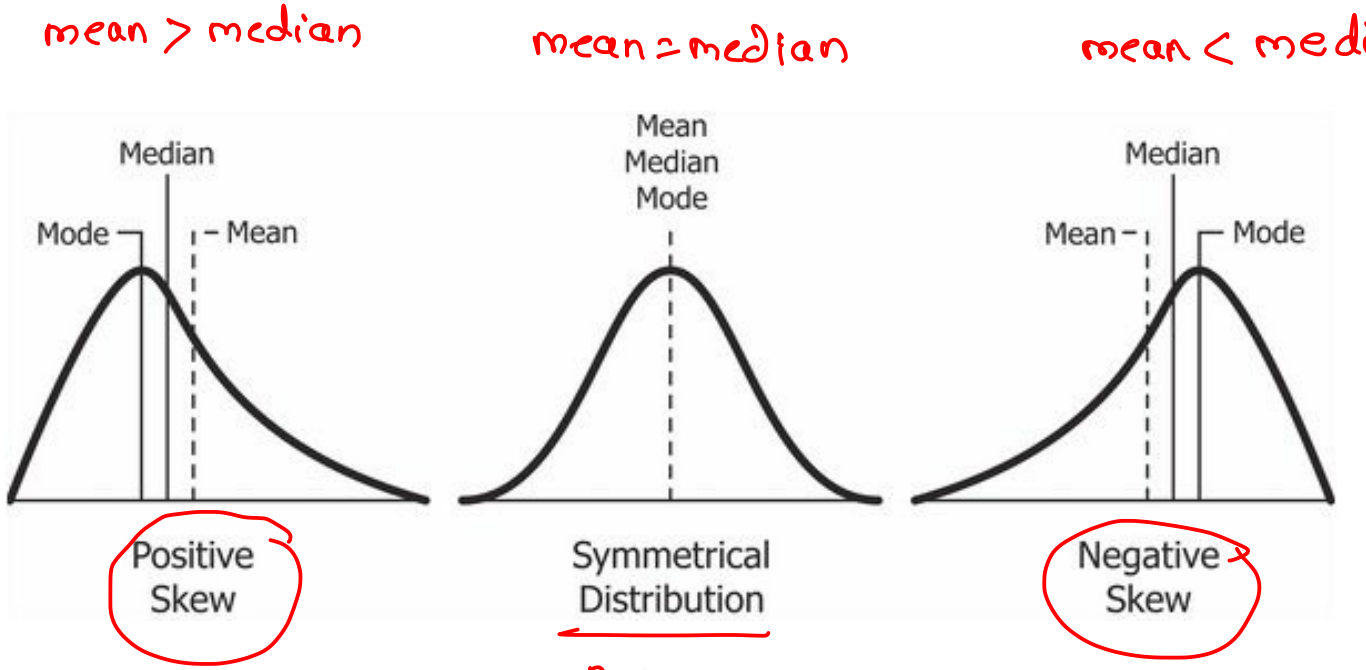
- Unlike discrete distributions, where the sum of all the bars equal to one, in a normal distribution the area under the curve equals to 1





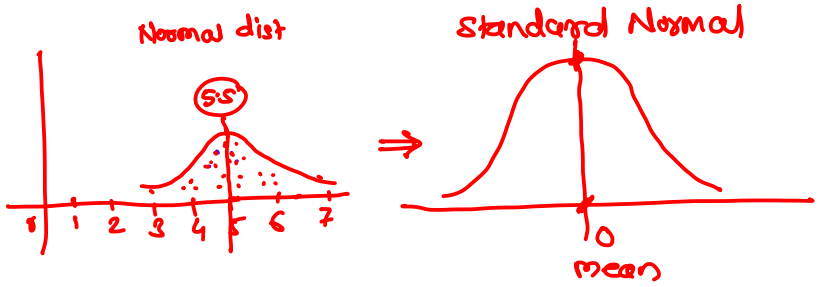
Normal Distribution

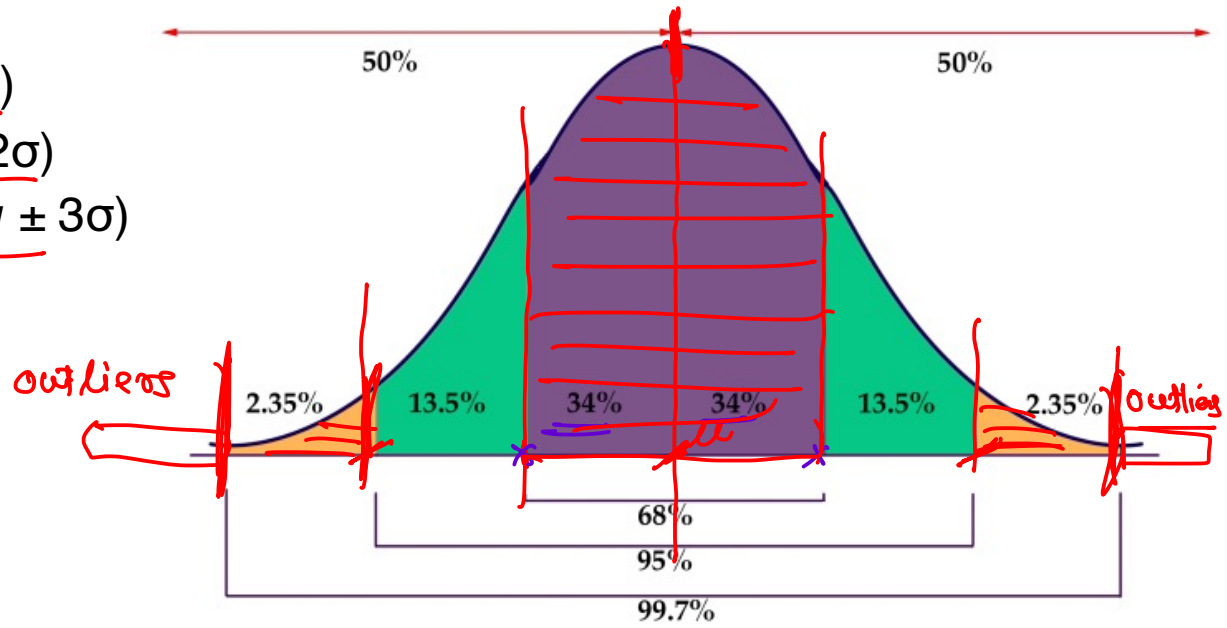
- Also called as Bell Curve or Gaussian Distribution
- Always symmetrical
- Asymmetrical curves display skew and are NOT normal



Standard Normal Distribution

- The standard normal distribution is a special case of the normal distribution
- It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one [$\mu = 0, \sigma = 1$]
- Empirical Rule 68-95-99.7
  - Often used for forecasting
  - For a normal distribution, almost all data falls within three standard deviations (σ) of the mean (μ)
  - The empirical rule shows that
    - 68% falls within the first standard deviation ($\mu \pm \sigma$)
    - 95% within the first two standard deviations ($\mu \pm 2\sigma$)
    - 99.7% within the first three standard deviations ($\mu \pm 3\sigma$)





Standardizing Normal Distribution

- We can take a normal distribution and standardize it to a standard normal distribution
- If we can model our data as a normal distribution we can convert the values in the normal distribution to a standard normal distribution to calculate percentile
- To calculate z-score use formula:

x : observation
μ : mean
σ : std devian

z-value

$$z = \frac{x - \mu}{\sigma}$$

- Using Z-table we can get the percentile of the value
- Since total area under the curve is 1, probabilities are bounded by 0 and 1

Central Limit Theorem

- The mean values from a group of samples will be normally distributed about the population mean, even if the population itself is not normally distributed
- That is, 95% of all sample means should fall within 2σ of the population mean
- To read more about CLT, please visit
  - https://en.wikipedia.org/wiki/Central_limit_theorem