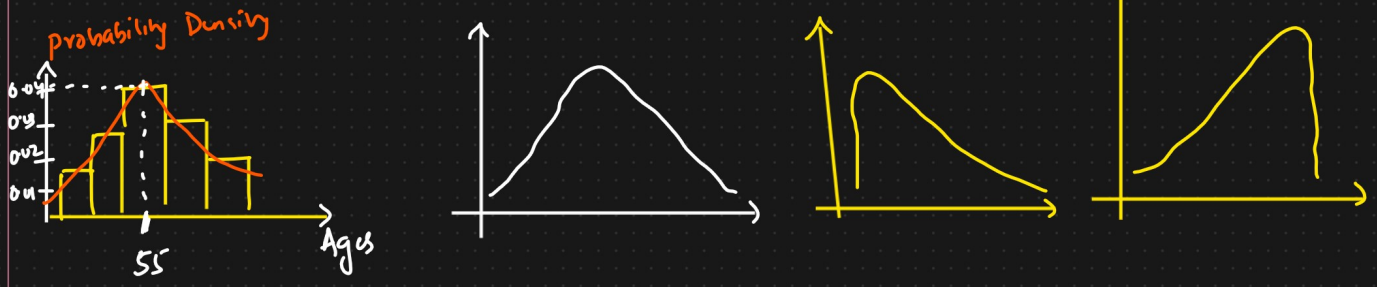
Probability Distribution Functions

Probability distribution functions describe how the probabilities are distributed over the values of a random variable.

Ages = { - - - - - - } ⇒ Continuous random variable



2 Main of probability distribution functions

① Probability Mass functions (PMF) : Used for discrete random variable.

② Probability Density functions (PDF) : Used for continuous random variable

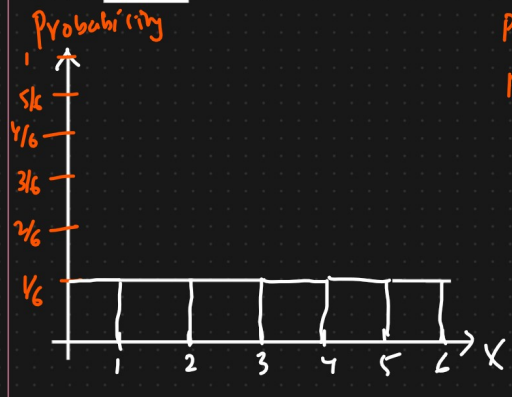
③ Cumulative Distribution function (cdf).

① Probability Mass Function [Discrete Random Variable]

Eg: Rolling a dice X = {1, 2, 3, 4, 5, 6} ⇒ Fair Dice

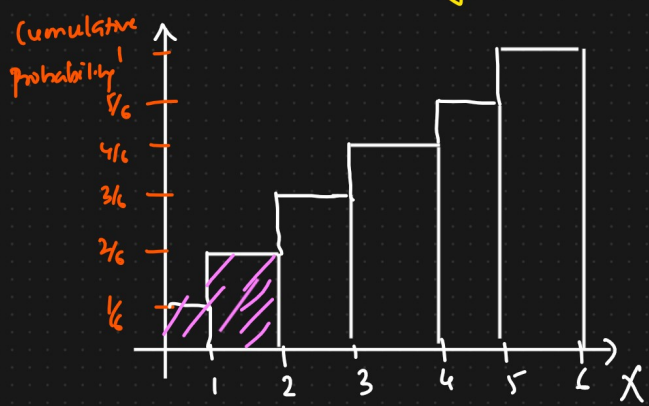
$Pr(1) = Pr(2) = Pr(3) = Pr(4) = Pr(5) = Pr(6) = 1/6$

PMF



$Pr(1) = \frac{1}{6}$
$Pr(2) = \frac{1}{6}$
$Pr(3) = \frac{1}{6}$
⋮
⇒

Cumulative Density Function (cdf)



$Pr(X \leq 2) = Pr(X=1) + Pr(X=2)$
$= 1/6 + 1/6 = 2/6 = 1/3$ //

$Pr(X \leq 6) = Pr(X=1) + Pr(X=2) + \dots + Pr(X=6)$
$= 1$ //

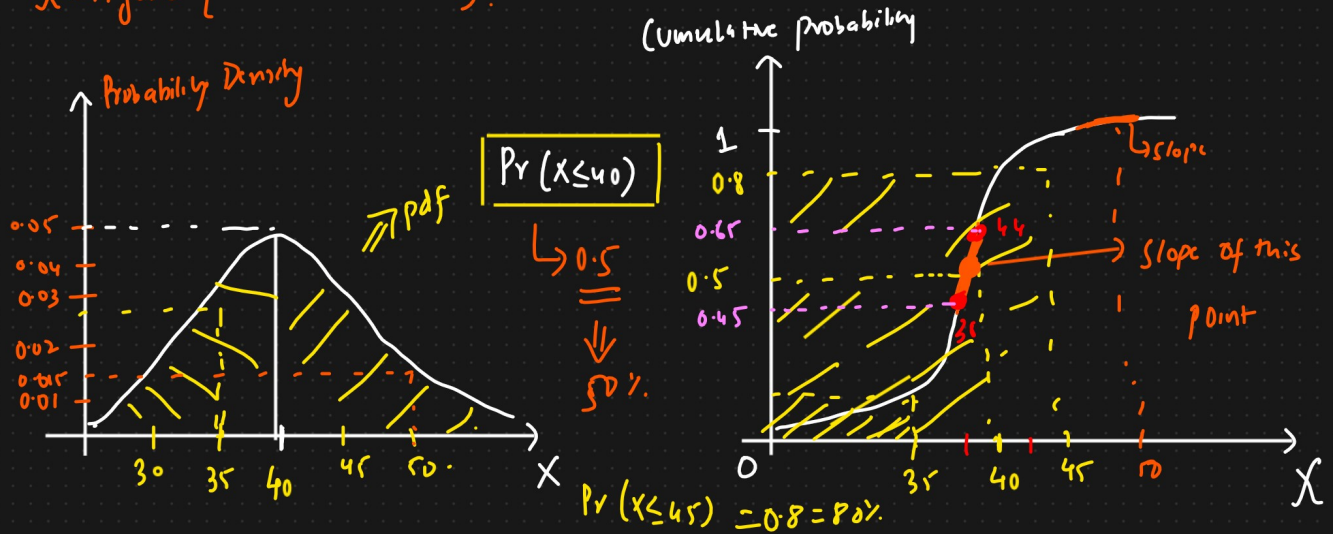
② Probability Density Function (pdf)

① Dismbution of Continuous Random Variable

① Area under the curve ✓

② Probability density ✓

$X = Ages = \{ - \; - \; - \; - \; - \}$



(x_1, y_1)

(x_2, y_2)

$\text{Slope} = \left\{ \frac{0.65 - 0.45}{44 - 38} \right\} \Rightarrow$ probability density

$\text{Slope} = \left[\frac{y_2 - y_1}{x_2 - x_1} \right]$ ⇐ Gradient ⇒ probability Density.

Probability Density = Gradient of Cumulative Density function

PDF Properties

① Non Negativity $f(x) \geq 0$ for all x

② The total area under the PDF curve is equal to 1

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

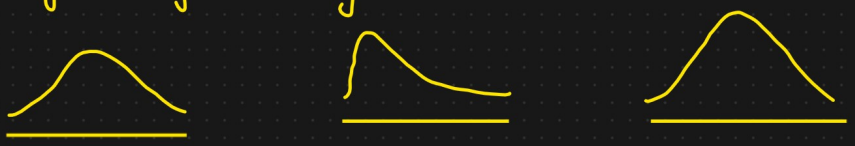
With respect to diffcent distribution

f(x) function is going to change

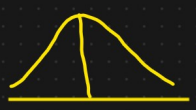
Different distribution types

Types of Probability Distribution [pdf, pmf, cdf]

Ages, Weight, Salary



DATASET ⇒ Distribution
⇓ ↑
SETS OF

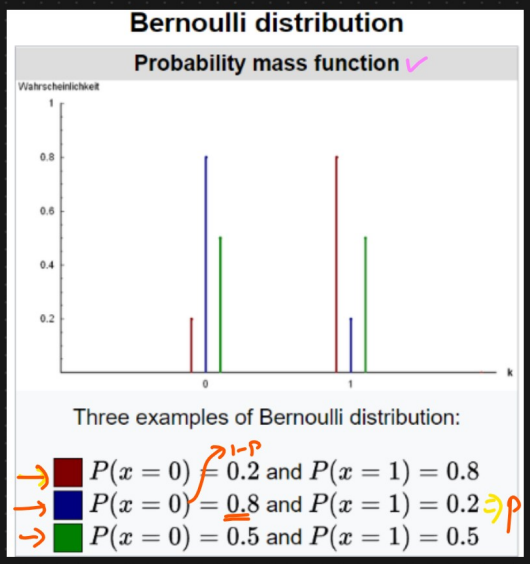
1. Bernoulli Distribution → Outcomes are binary (pmf) ⇒ Discrete Random Variable
2. Binomial Distribution → (pmf)
3. Normal/Gaussian Distribution → (pdf) ⇒  ⇒ Assumptions.
4. Poisson Distribution (pmf)
5. Log Normal Distribution (pdf)
6. Uniform Distribution (pmf)

Dataset → House price prediction Dataset- [EDA, FE] ⇒ DATA ANALYST, DATA SCIENTIST

Size of the house	No. of Rooms	Location	Floor	Sea Side	Price.
↓	↓	↓	↓	↓	
Continuous random Variable	{Discrete}		{Discrete}	{0 & 1}	Continuous
pdf.			pmf	pmf	pdf

Bernoulli Distribution

Definition: The Bernoulli distribution is the simplest discrete probability distribution. It represents the probability distribution of a random variable that has exactly two possible outcomes: success (**with probability p**) and failure (**with probability 1−p**). It is used to model binary outcomes, such as a coin flip or a yes/no question.



① Discrete Random Variable (pmf)

② Outcomes are Binary

Eg: ① Tossing a coin {H, T}

$Pr(X = H) = 0.5 = p$

$Pr(X = T) = 1 - 0.5 = 0.5 = q.$

P, q

$q = (1-p)$

② Whether the person will Pass/Fail

$Pr(X = Pass) = 0.4$

$Pr(X = Fail) = 1 - 0.4 = 0.6.$

Parameters

$0 \leq p \leq 1$

$q = 1 - p$

$K = \{0, 1\}$ ⇒ 2 outcomes.

$Pr(Success) \Rightarrow K = 1$

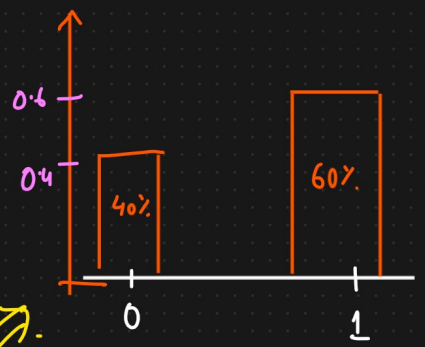
$Pr(Fail) \Rightarrow K = 0$

PMF : Company has launched a new Smartphone 'A'

K

(1) Use = 60% ⇒ p

(0) Not use = 40% ⇒ $q = 1 - p$



⇐ Distribution

$PMF = p^K * (1-p)^{1-K}$

if K = 1

$$Pr(K=1) = p^1 (1-p)^{1-1} \Rightarrow p \;//$$

$$Pr(K=0) = p^0 * (1-p)^1 \Rightarrow (1-p) = q \;//$$

Simplified

$$pmf \begin{cases} q = 1-p & \text{if } K=0 \\ p & \text{if } K=1 \end{cases}$$

ⓔ Mean of Bernoulli Distribution

$$E(x) = \sum_{k=0}^{1} K \cdot p(K) \qquad K=\{0,1\}.$$

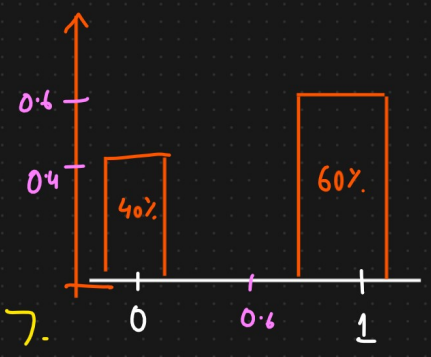
$$= 0 \times 0.40 + 1 * 0.60 \qquad p = 0.6$$

$$= 0 + 0.60 \qquad q = 0.4$$

$$= 0.60 \Rightarrow p \Leftarrow \;//$$

$$Pr(1) = 0.6)$$

$$Pr(0) = 0.4$$



ⓕ Median of Bernoulli Distribution

$$\text{Median} \begin{cases} 0 & \text{if } p < 1/2. \\ \rightarrow [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$$

$$\begin{cases} \text{median} = 0 & \text{if } q > p \\ \text{median} = 0.5 & \text{if } q = p \\ \text{median} = 1 & \text{if } q < p. \end{cases}$$

ⓖ Mode

$p > q \Rightarrow p$ will be the mode

else q will be the mode.

④ Variance

$K = 0$ and 1 $\quad Pr(k=0) = 0.4 \Rightarrow q$ $\quad Pr(k=1) = 0.6 \Rightarrow p$

$$\sigma^2 = 0.40 * (0 - 0.6)^2 + 0.6 (1 - 0.6)^2$$

$$= 0.40 * 0.36 + 0.6 (0.16)$$

$$\sigma^2 = 0.24 \Rightarrow Pr(k=0) * Pr(k=1)$$

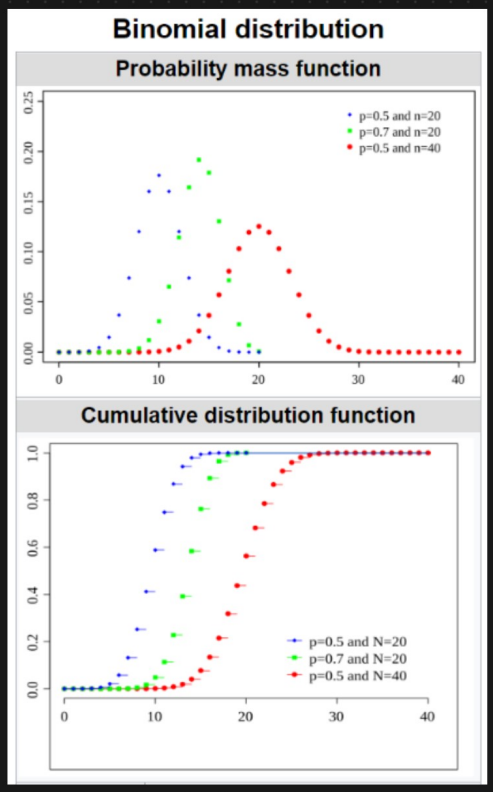
$$q * p$$

$$\sigma^2 = pq$$

$$\sigma = \sqrt{pq}$$

Binomial Distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability q = 1-p). A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.



(1) Discrete Random Variable

(2) Every outcome of the experiment is binary

(3) These experiments are performed for n trials

Eg: Tossing a coin 10 times $\boxed{n=10}$

↓

$\{H, T\}$

Notation : $B(n, p)$

Parameters : $n \in \{0, 1, 2, \cdots\} \Rightarrow$ no. of trails or experiment

$p \in [0, 1] \rightarrow$ Success probability for each trial

$q = 1 - p$

Support : $k \in \{0, 1, 2, 3 \cdots n\} \Rightarrow$ Number of successes

PMF :

$$Pr(k, n, p) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$$

for $k = 0, 1, 2, \cdots n$ where

$$\boxed{{}^{n}C_{k} = \frac{n!}{k!(n-k)!}} \Rightarrow \text{Binomial Coefficient.}$$

Mean : $n \cdot p$

Variance : npq.

σ : $\sqrt{npq}$

Eg : Coin flip

No. of trial $(n) = 5$

Probability of success $(p) = 0.5$

No. of Success (K) = Varies from 0 to 5

① What is the probability of getting exactly 3 heads in 5 flips?

$n = 5 \quad K = 3$

$$P_Y(X = 3) = {}^{5}C_{3} (0.5)^{3} (1 - 0.5)^{5-3} = 0.3125.$$

Example : Quality Control

Scenario : Inspecting 10 items in a factory where each item has a 10% chance of being defective

① No. of Trials $(n) = 10$

② Probability of Success $(p) = 0.1$ (defective item)

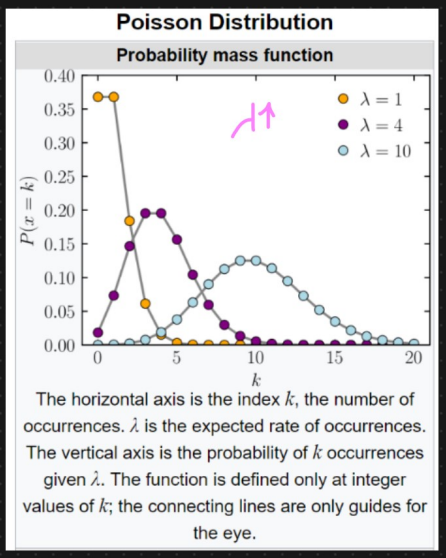
③ No. of Successes (K) = Varies from 0 to 10

Question: What is the probability of finding exactly 2 defective items in a sample of 10?

$$Pr(X=2) = {}^{10}C_2 (0.1)^2 (1-0.1)^{10-2} \approx 0.1937 //.$$

Poisson Distribution

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known constant mean rate and independently of the time since the last event.



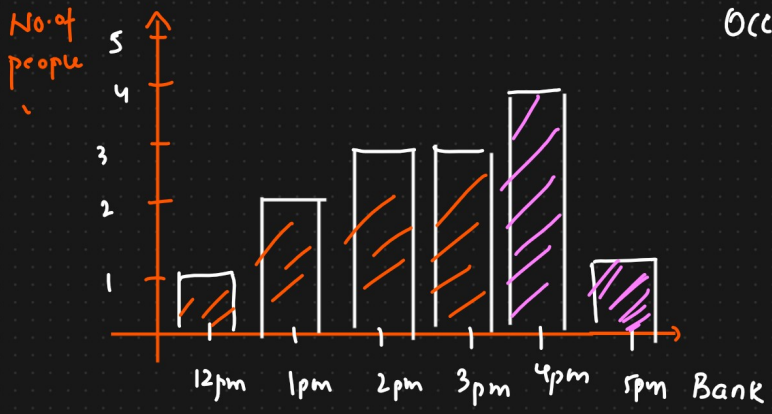
The horizontal axis is the index k, the number of occurrences. λ is the expected rate of occurrences. The vertical axis is the probability of k occurrences given λ. The function is defined only at integer values of k; the connecting lines are only guides for the eye.

① Discrete random variable (pmf)

② Describes the numbers of events occuring in a fixed time intervals

Eg:- No. of people visiting hospital every hour
No. of people visiting banks every hour

$\lambda = 3$ ⇒ Expected no. of events occuring at every time interval



PMF

$\lambda = 3$

$$P(X=5) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

$$= \frac{e^{-3} 3^{5}}{5!} = 0.101 = 10.1\%$$

$Pr(X \leq 3)$

$Pr(X=4) + Pr(X=5) =$ final ⇒ probability

Mean of Poisson Distribution

λ = Expected No. of events occur at every time interval

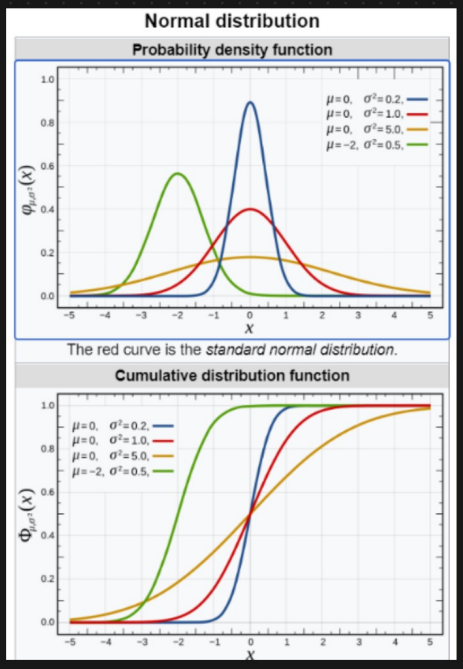
t = Time interval

$$\text{Mean} = E(x) = \mu = \lambda * t$$

Variance

Normal / Gaussian Distribution

In probability theory and statistics, a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable



① Continuous random variable (pdf)

$X = \{ - - - - \}$

$\sigma^2 \uparrow\uparrow$

Spread $\uparrow\uparrow$

⇒ Symmetric distribution

M = median = mode

Eg: Weights of students in a class {Doctors}

Heights of students in a class {Doctors}

IRIS DATASET → Petal length, Sepal Length

Petal Width Sepal Width

Notation $N(\mu, \sigma^2)$

Researchers

Parameters : $\mu \in R$ = mean

$\sigma^2 \in R > 0$ = Variance

$x \in R$.

$$PDF = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

Mean of Normal / Gaussian

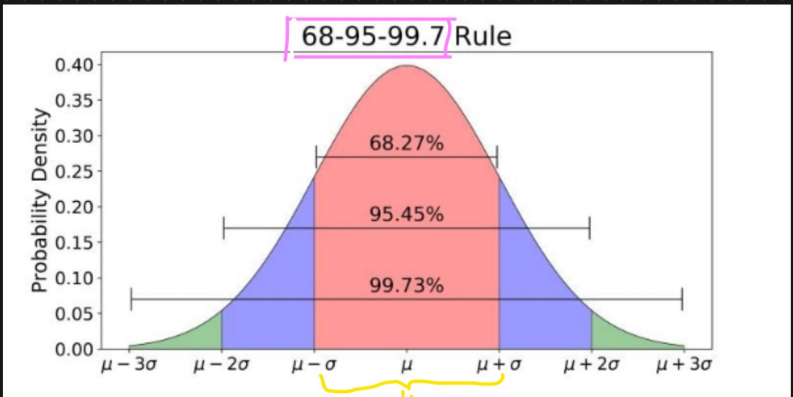
$$\mu = \sum_{i=1}^{n} \frac{x_i}{n}$$

Variance

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

$\sigma = \sqrt{\text{variance}}$

Empirical Rule of Normal / Gaussian Distribution



QQ plot

$X = \{ - \quad - \quad - \quad - \quad - \quad \cdots \}$ => Normal / Gaussian Distribution

- Probability

$$Pr(\mu-\sigma \leq X \leq \mu+\sigma) \approx 68\%$$

$$Pr(\mu-2\sigma \leq X \leq \mu+2\sigma) \approx 95\%$$

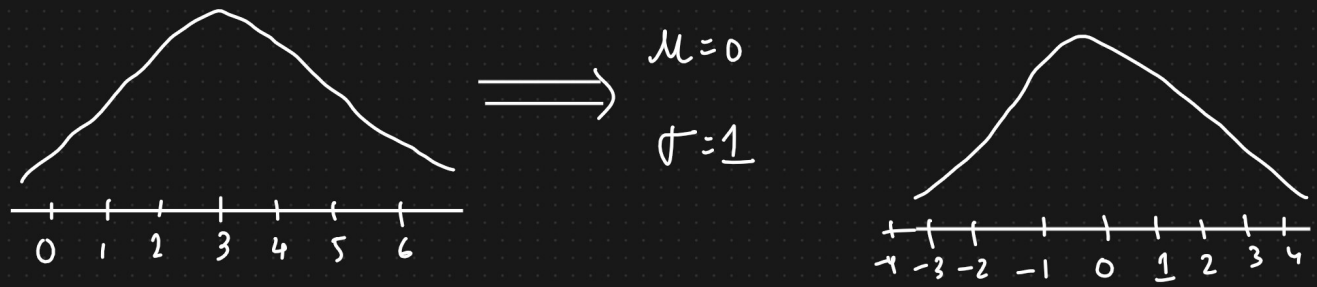
$$Pr(\mu-3\sigma \leq X \leq \mu+3\sigma) \approx 99.7\%$$

① Standard Normal Distribution

Z-score

$X = \{1, 2, 3, 4, 5\}$ $\quad \mu = 3$

$\sigma = 1.414 \approx 1$

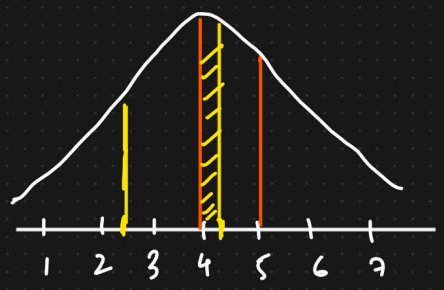


$X = \{1, 2, 3, 4, 5\}$ $\quad$ Z-score $= \dfrac{x_i - \mu}{\sigma}$ $\quad Y = \{-2, -1, 0, 1, 2\}$

① $\dfrac{1-3}{1} = -2$ $\quad$ ③ $\dfrac{3-3}{1} = 0$

② $\dfrac{2-3}{1} = -1$ $\quad$ ④ $\dfrac{4-3}{1} = 1$

$$X \approx SND(\mu = 0, \sigma = 1)$$



$\mu = 4$

$\sigma = 1$

Q) How many standard deviation 4.25 is away from the mean?

$x_i = 4.25$

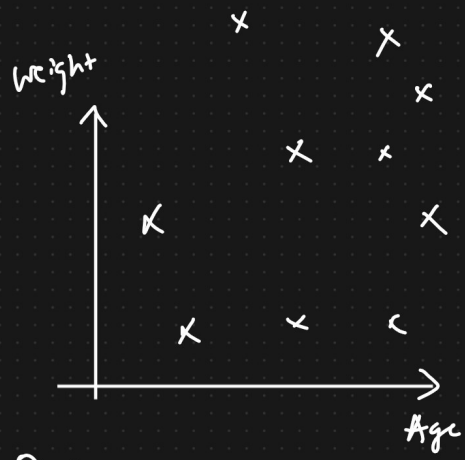
$$Z\text{-score} = \frac{4.25 - 4}{1} = 0.25$$

$x_i = 2.5$

$$Z\text{-score} = \frac{2.5 - 4}{1} = -1.5$$

Eg: Dataset

Age (years)	Weight (kg)	Height (cms)	Salary (INR)
24	70	175	40K
25	60	160	50K
25	55	180	60K
22	40	130	30K
30	30	175	20K
31	25	180	70K
⇓	⇓	⇓	⇓



① Clustering Algorithms
② Linear Regression
③ Logistic Regression

Standardization ⇒ ML Models

$$Z\text{-score} = \frac{x_i - \mu_{age}}{\sigma} \qquad \frac{x_i - \mu_{weight}}{\sigma}$$

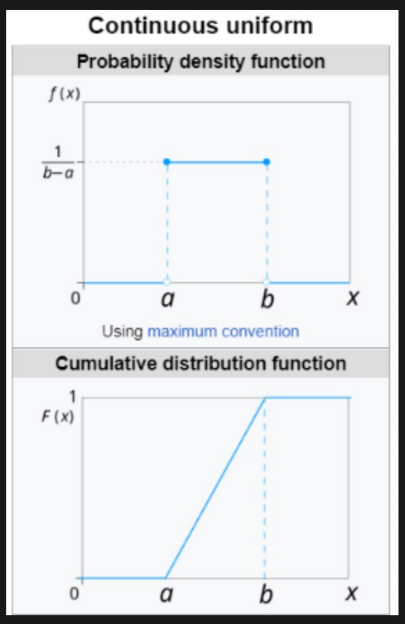
Uniform Distribution

① Continuous Uniform Distribution (pdf)

② Discrete Uniform Distribution (pmf)

① Continuous Uniform Distribution [Continuous Random Variable]

In probability theory and statistics, the continuous uniform distributions or rectangular distributions are a family of symmetric probability distributions. Such a distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds. The bounds are defined by the parameters, a and b which are the minimum and maximum values.



Notation : $U(a,b)$

Parameters : $-\infty < a < b < \infty$

$$\text{Pdf} = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Cdf} = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$$

Mean $= \frac{1}{2}(a+b)$ //

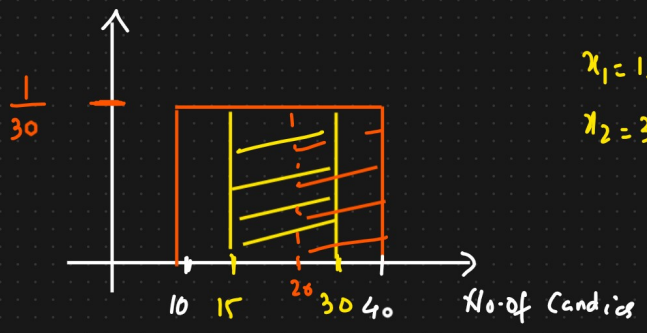
Median $= \frac{1}{2}(a+b)$ //

Variance $= \frac{1}{12}(b-a)^2$ =

Eg : The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 candies and a minimum of 10

i) Probability of daily sales to fall between 15 and 30?

Ans)



$x_1 = 15$

$x_2 = 30$

$$Pr(15 \le X \le 30) = (x_2 - x_1) * \frac{1}{b-a}$$

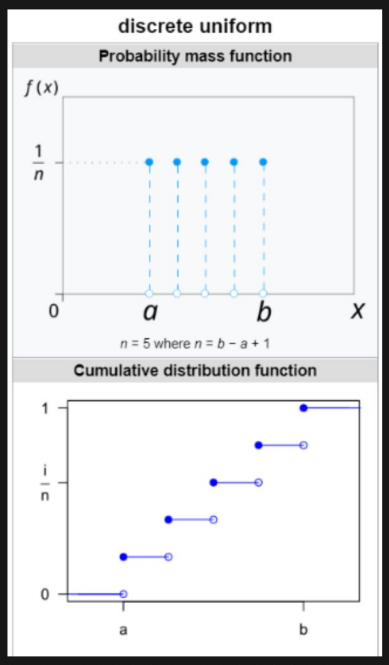
$$= (30-15) * \frac{1}{30}$$

$$= 0.5 //$$

$$Pr(X \ge 20) = (40-20) * \frac{1}{30} = 0.66 = 66\%.$$

② Discrete Uniform Distribution

In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution wherein a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n. Another way of saying "discrete uniform distribution" would be "a known, finite number of outcomes equally likely to happen".



④ Discrete Random Variable

⑤ pmf

Eg: Rolling a dice ⇒ Fair dice, a ↓ 1, b ↓ 6, {1, 2, 3, 4, 5, 6}

Pr(1) = 1/6
Pr(2) = 1/6
Pr(3) = 1/6
⋮
= 1/6

1/n ⇒ n = b − a + 1 = 6 − 1 + 1 = 6

1/6

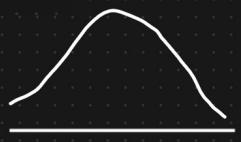
Notation $U(a,b)$

Parameters a, b where $b \geq a$

PMF $1/n$

Mean, Median ⌒ $\frac{a+b}{2}$

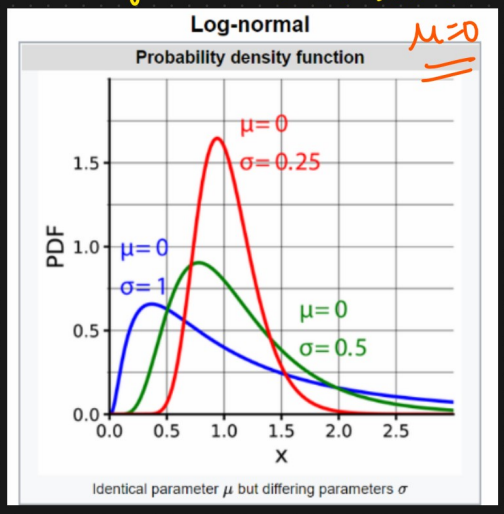
① Log Normal Distribution



Empirical Rule

In probability theory, a log-normal (or lognormal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.Equivalently, if Y has a normal distribution, then the exponential function of Y, X = exp(Y), has a log-normal distribution

Right Skewed Distribution



$\mu = 0$

$X \approx$ Log Normal Distribution(μ, σ)

⇓

$Y \approx \ln(X)$ = Normal Distribution

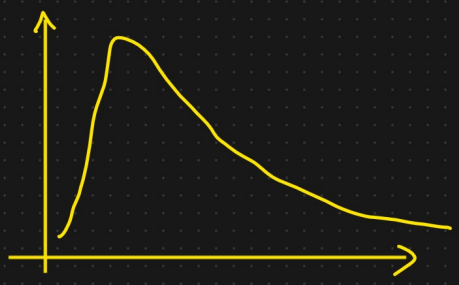
⇓

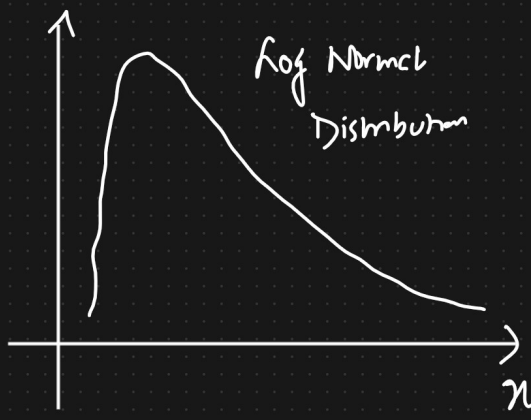
Natural log

$[\log_e]$

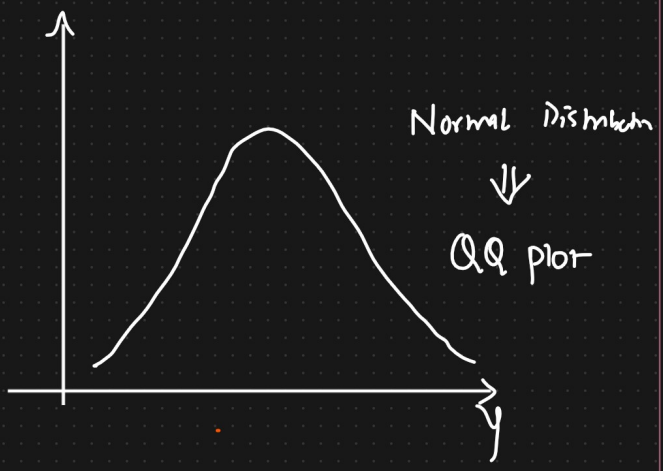
⇓

$X \approx \exp(Y) \Rightarrow$ Log Normally Distributed

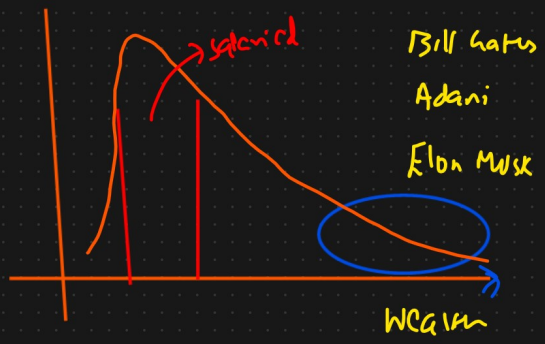




$\ln(x)$ ⟹ $\exp(y)$



Eg :



① Wealth distribution of the world

② Discussion Forum → length of the comments

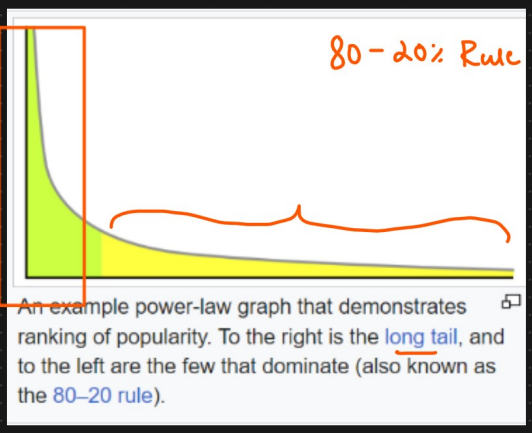
③ length of chess games

④ Dwell time on online articles (joke, news)

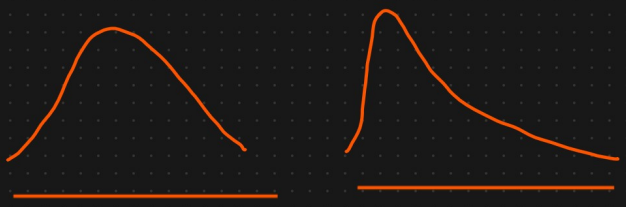
⑤ Salaries of employees in a company.

① Power Law Distribution

In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).



Eg: IPL

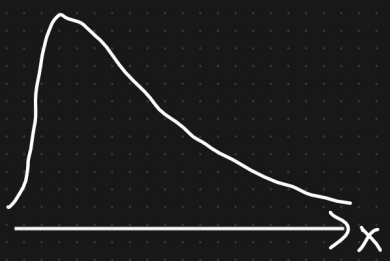
① 20% of Team is responcible for winning 80% match

② 80% of wealth are distributed with 20% of the total Population

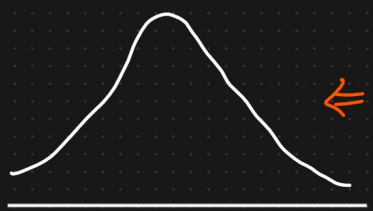
③ 80% of the total Oil is with 20% of the nation

④ Frequencies of words in most languages.

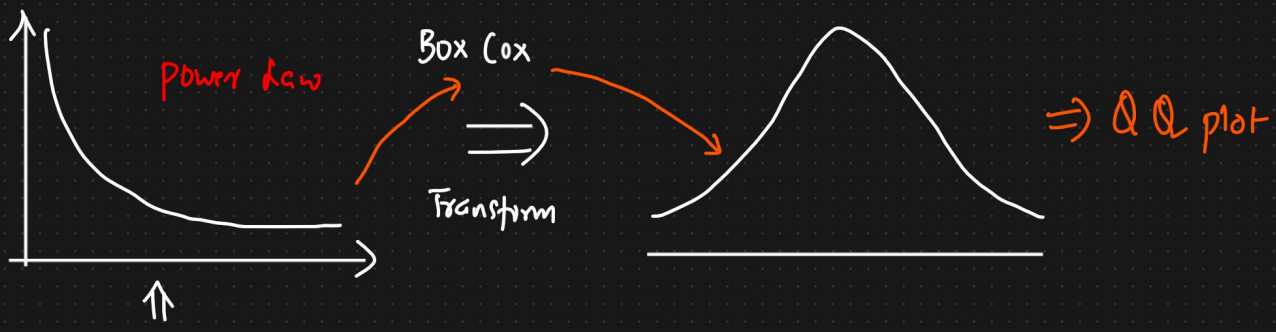
⑤ 20% of the major defects fixes the 80% of upcoming defects in a S/w product



ln(x) ⟹

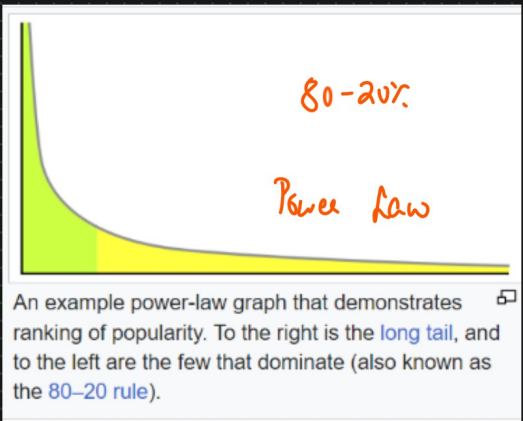


⇐ Normal Distribution



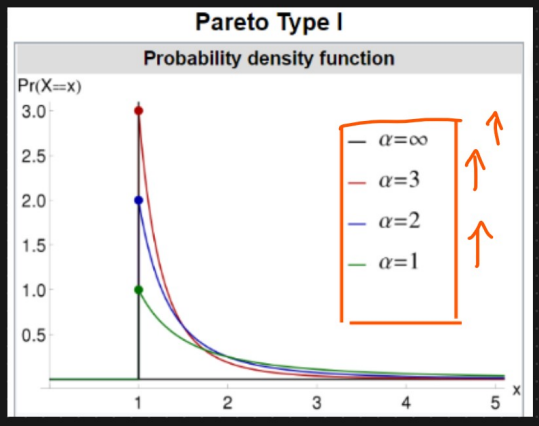
Pareto Distribution [80-20%]

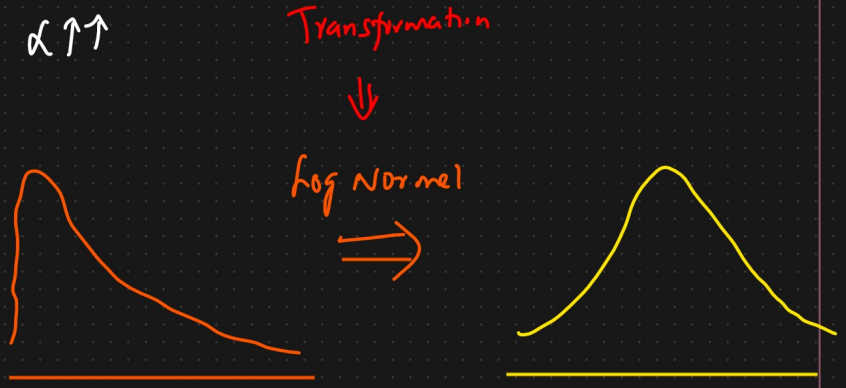
① Pareto Distribution [NON Gaussian Distribution]

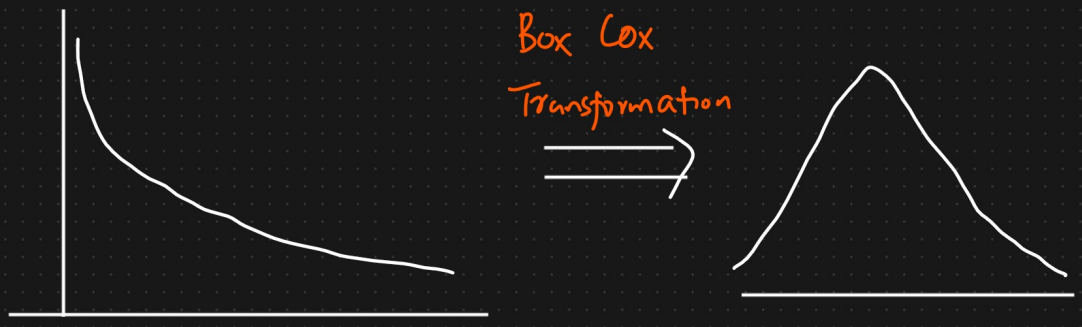


An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

The Pareto distribution, named after the Italian civil engineer, economist, and sociologist Vilfredo Pareto is a power-law probability distribution that is used in description of social, quality control, scientific, geophysical, actuarial, and many other types of observable phenomena; the principle originally applied to describing the distribution of wealth in a society, fitting the trend that a large portion of wealth is held by a small fraction of the population.







Eg: IT Industry

① 80% of the entire project is done by 20% of the team

② 80% of defects can be solved if we save 20% of defects

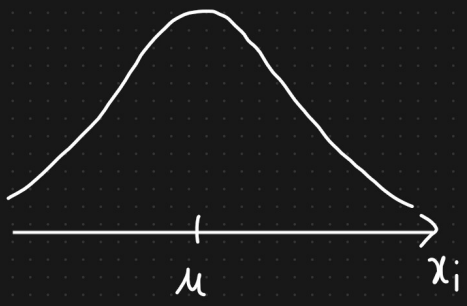
① Central Limit Theorem

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

① $X \sim N(\mu, \sigma)$

Population DATA



n = sample size ⇒ any value

$S_1 = \{x_1, x_2, x_4 \cdots x_n\} = \bar{x}_1$

$S_2 = \{x_2, x_3 \cdots x_n\} = \bar{x}_2$

S_3 $\bar{x}_3$

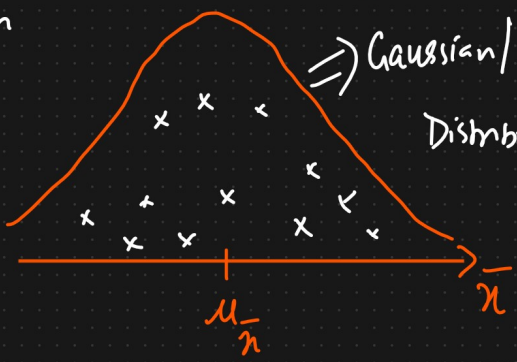
S_4 ⋮

⋮ $\bar{x}_m$

S_m

⇓ CLT ↘

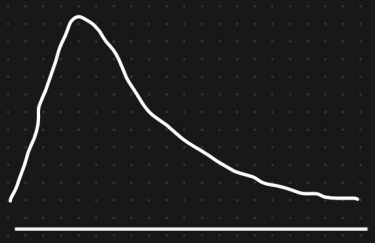
Sampling distribution of the mean



⇒ Gaussian/Normal Distribution

② $X \not\sim N(\mu, \sigma)$

⇒ $n \geq 30$ ⇒ sample size



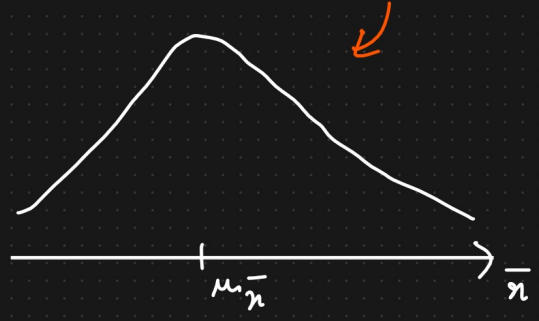
S_1 $\bar{x}_1$

S_2 $\bar{x}_2$

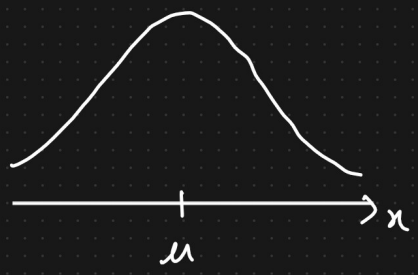
⋮

$\bar{x}_m$

⇓ CLT



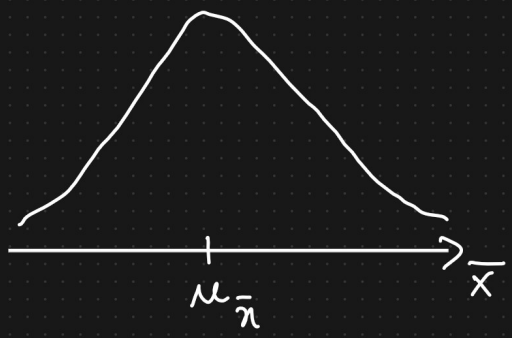
① Normal Distribution



$$X \approx N(\mu, \sigma)$$

⇓

Sampling Distribution of mean



σ = population std

μ = population mean

n: sample size

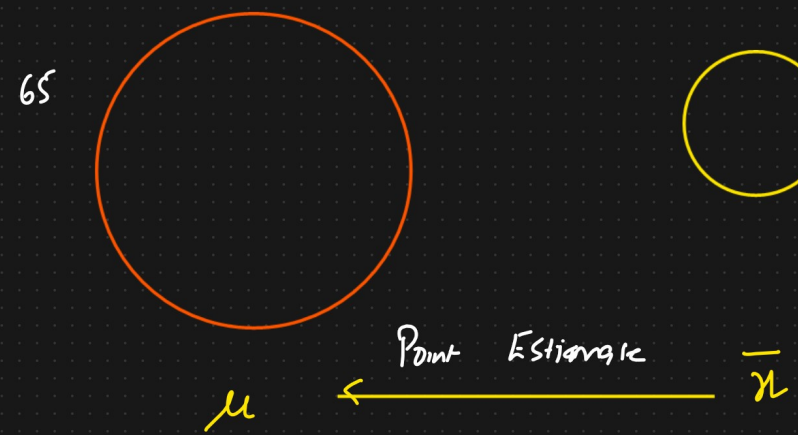
$$X \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Estimate: It is a specified observed numerical value used to estimate an unknown population parameter

Types

① **Point Estimate**: Single numerical value used to estimate an unknown population parameter

Eg: Sample mean is a point estimate of a population mean



② **Interval Estimate**: Range of values used to estimate the unknown population parameter

55 - 65 ⇒ Sample Mean