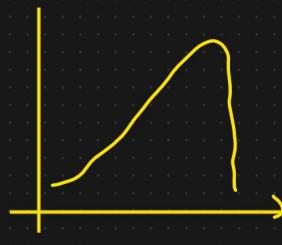
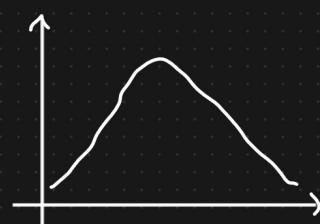
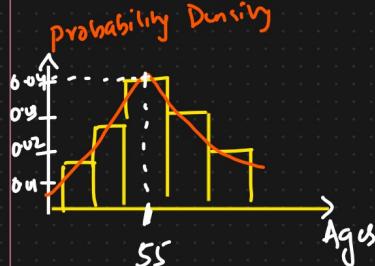


# Probability Distribution Functions

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

$A_{\text{Age}} = \{ \dots \} \Rightarrow$  continuous random variable

probability Density



2 Main of probability distribution functions

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

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

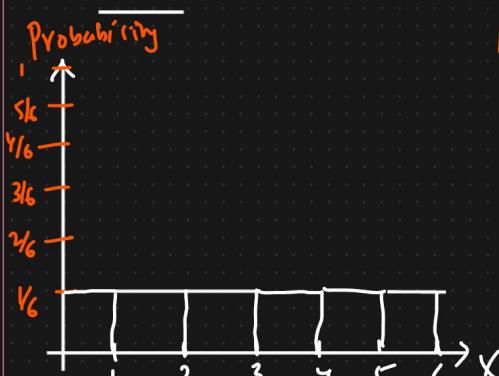
③ Cumulative Distribution function (cdf).

① Probability Mass Function [Discrete Random Variable]

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

$$Pr(1) = Pr(2) = Pr(3) = Pr(4) = Pr(5) = Pr(6) = \frac{1}{6}$$

PMF



$$Pr(1) = \frac{1}{6}$$

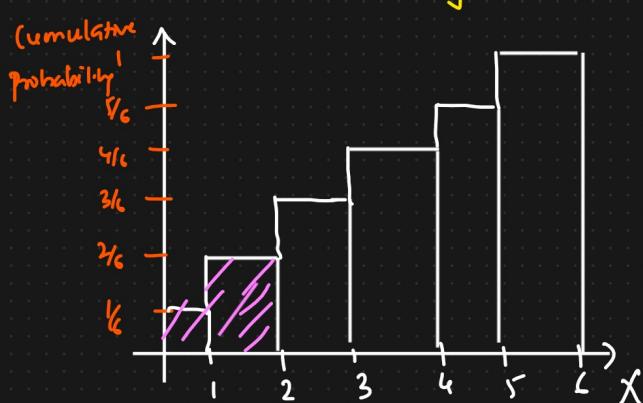
$$Pr(2) = \frac{1}{6}$$

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

$$\vdots$$

$\Rightarrow$

Cumulative Density Function (cdf)



$$Pr(X \leq 2) = Pr(X=1) + Pr(X=2)$$

$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

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

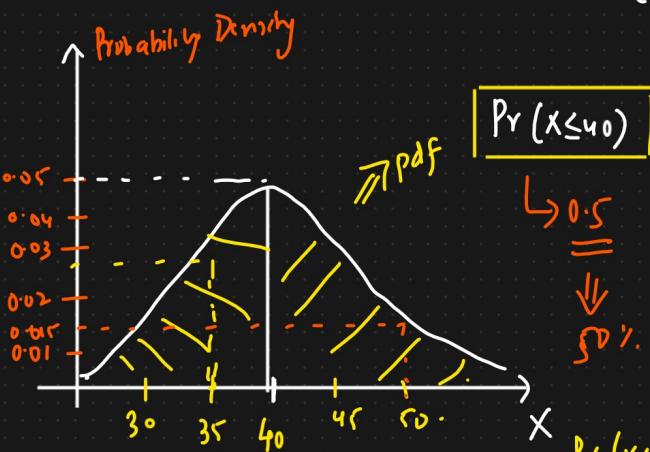
$$+ \dots + Pr(X=6) = 1$$

## ② Probability Density Function (pdf)

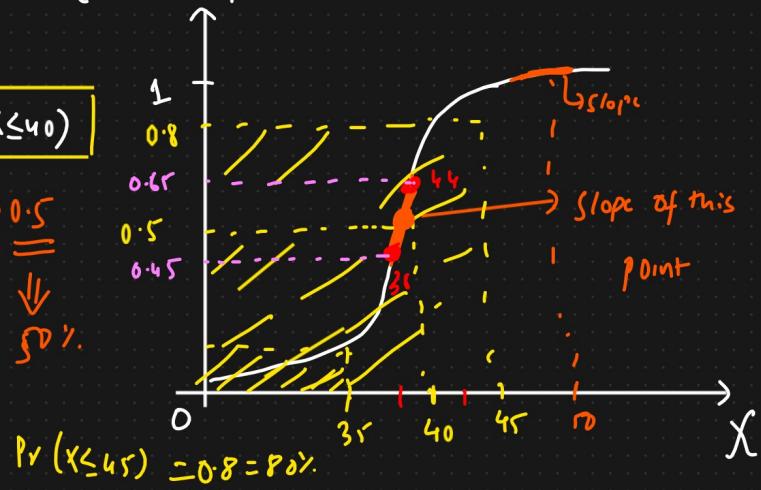
① Distribution of Continuous Random Variable

① Area under the curve ✓  
② Probability density ✓.

$$X = \text{Age} = \{ \dots \}$$



(Cumulative probability)



$$\text{Slope}_{\text{c}} = \left[ \frac{y_2 - y_1}{x_2 - x_1} \right] \Leftrightarrow \text{Gradient} \Rightarrow \text{probability density}$$

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 different distribution

$f(x)$  function is going to change



## Different distribution types

### Types of Probability Distribution

[pdf, pmf, cdf]

Age, Weight, Salary



DATASET  $\Rightarrow$  Distribution  
SETS OF  $\Downarrow \Uparrow$

- ① Bernoulli Distribution  $\rightarrow$  Outcomes are binary (pmf)  $\Rightarrow$  Discrete Random Variable
- ② Binomial Distribution  $\rightarrow$  (pmf)
- ③ Normal/Gaussian Distribution  $\rightarrow$  (pdf)  $\Rightarrow$  Assumptions.
- ④ Poisson Distribution (pmf)
- ⑤ Log Normal Distribution (pdf)
- ⑥ Uniform Distribution (pmf)

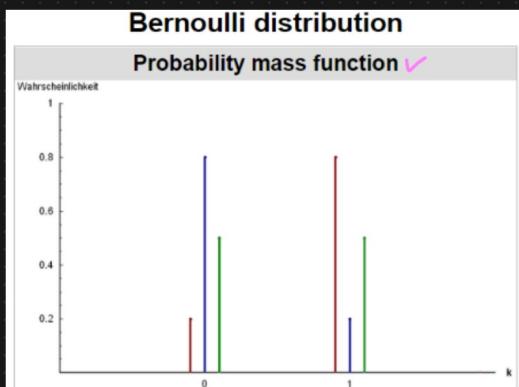
Dataset  $\rightarrow$  House price prediction Dataset

[EDA, FE]  $\Rightarrow$  DATA ANALYST  
DATA SCIENTIST

Size of the house	No. of Rooms	location	Floor	Sca Side	Price.
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	
{Continuous Random Variable}	{Discrete}		{Discrete}	{ <u>0 &amp; 1</u> }	(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.



Three examples of Bernoulli distribution:

- $P(x=0) = 0.2$  and  $P(x=1) = 0.8$
- $P(x=0) = 0.8$  and  $P(x=1) = 0.2$
- $P(x=0) = 0.5$  and  $P(x=1) = 0.5$

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

$$\Pr(X=\text{Pass}) = 0.4$$

$$\Pr(X=\text{Fail}) = 1 - 0.4 = 0.6$$

## Parameters

$$0 \leq p \leq 1$$

$$q = 1-p$$

$$K = \{0, 1\} \Rightarrow 2 \text{ outcomes.}$$

$$\Pr(\text{Success}) \Rightarrow K=1$$

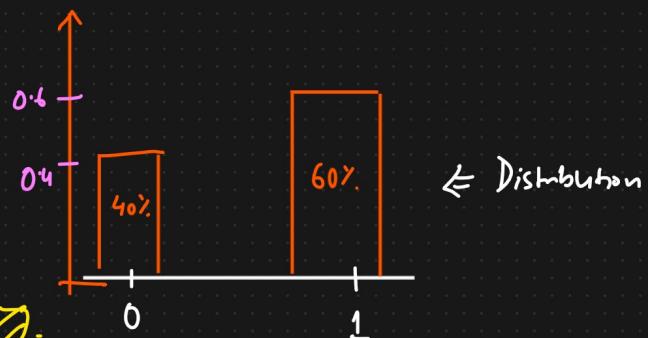
$$\Pr(\text{Fail}) \Rightarrow K=0$$

PMF  $\hat{=}$  Company has launched a new Smartphone 'A'

K

$$(1) \text{ Use } = 60\% \Rightarrow p$$

$$(0) \text{ Not Use } = 40\% \Rightarrow q = 1-p$$



$$\text{PMF} = P^K \cdot (1-p)^{1-k}$$

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

$$\text{pmf} \quad \left\{ \begin{array}{ll} q = 1-p & \text{if } K=0 \\ p & \text{if } K=1 \end{array} \right.$$

#### ④ Mean of Bernoulli Distribution

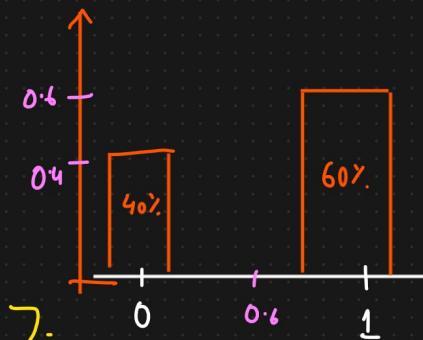
$$E(X) = \sum_{k=0}^1 k \cdot p(k) \quad K = \{0, 1\}$$

$$= 0 \times 0.40 + 1 \times 0.60 \quad p = 0.6$$

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

$$= 0.60 \Rightarrow p \leftarrow // \quad \Pr(1) = 0.6$$

$$\Pr(0) = 0.4$$



#### ④ Median of Bernoulli Distribution

$$\text{Median} \quad \left\{ \begin{array}{ll} 0 & \text{if } p < \frac{1}{2} \\ [0, 1] & \text{if } p = \frac{1}{2} \\ 1 & \text{if } p > \frac{1}{2} \end{array} \right.$$

$$\left\{ \begin{array}{ll} \text{median} = 0 & \text{if } q > p \\ \text{median} = 0.5 & \text{if } q = p \\ \text{median} = 1 & \text{if } q < p. \end{array} \right.$$

#### ④ Mode

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

④ Variance  $K=0 \text{ and } 1$   $P_{r(K=0)} = 0.4 \Rightarrow q, P_{r(K=1)} = 0.6 \Rightarrow p$

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

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

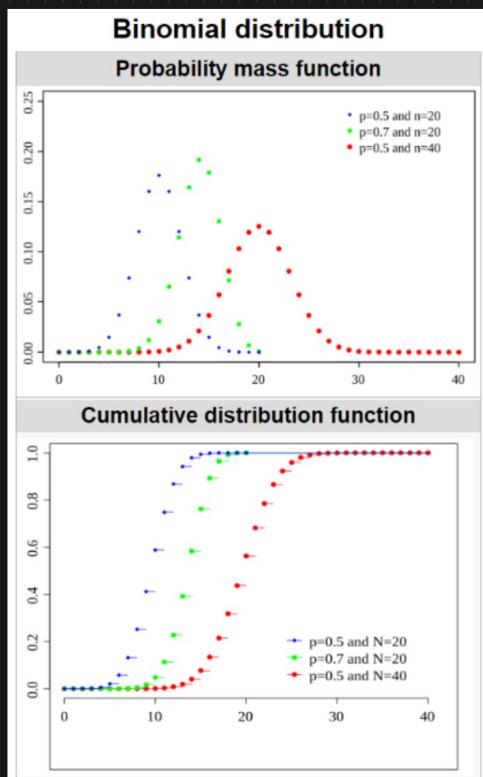
$$\sigma^2 = 0.24 \Rightarrow P_{r(K=0)} \neq P_{r(K=1)}$$

$$q \neq p$$

$$\begin{aligned} \sigma^2 &= pq \\ \sigma &= \sqrt{pq} \end{aligned}$$

## 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       $\boxed{n=10}$   
 $\downarrow$   
 $\{H, T\}$

Notation :  $B(n, p)$

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

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

$$q = 1 - p$$

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

PMF :

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

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

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \Rightarrow \text{Binomial Coefficients.}$$

$$\left. \begin{array}{l} \text{Mean : } np \\ \text{Variance : } npq \\ \sigma : \sqrt{npq} \end{array} \right\}$$

Eg: Coin flip

No. of trials ( $n$ ) = 5

Probability of success ( $p$ ) = 0.5

No. of Success ( $k$ ) = varies from 0 to 5

④ What is the probability of getting exactly 3 heads in 5 flips?

$n=5$     $K=3$

$$Pr(X=3) = \binom{5}{3} (0.5)^3 (1-0.5)^{5-3} = 0.3125$$

Example: Quality Control

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

④ No. of Trials ( $n$ ) = 10

④ Probability of Success ( $p$ ) = 0.1 (defective item)

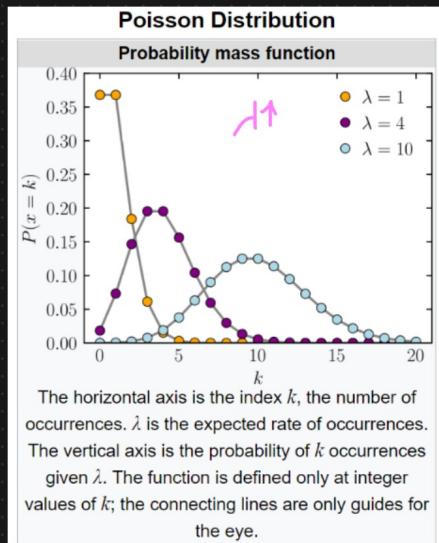
④ No. of Successes ( $k$ ) = varies from 0 to 10

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.

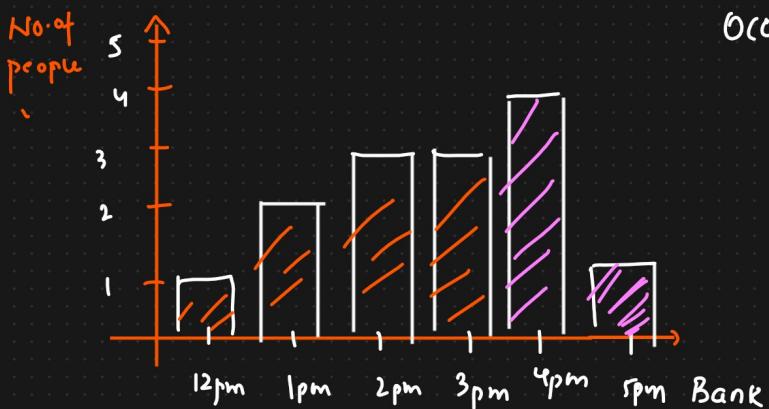


① Discrete random variable (pmf)

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

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

$\lambda = 3 \Rightarrow$  Expected no. of events  
occurring at every time interval



PMF

$$\boxed{\lambda = 3}$$

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

$$Pr(X \leq 3)$$

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

$$Pr(X=4) + Pr(X=5) = \text{final} \Rightarrow \text{probability}$$

## Mean of Poisson Distribution

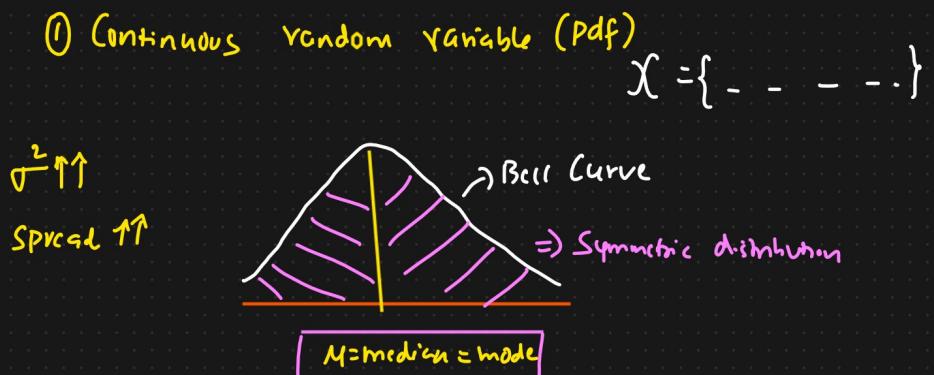
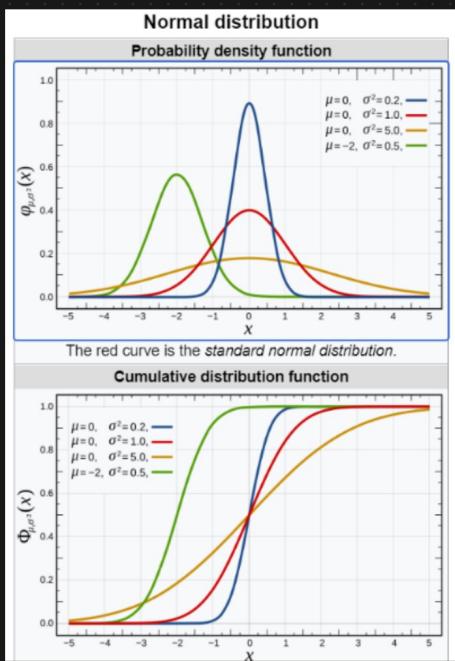
$$\text{Mean} = E(x) = M = \lambda * t$$

Variance

$\lambda$  = Expected No. of events occur at  
every time interval  
 $t$  = Time intervals

# 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



Eg: Weights of students in a class {Doctors}

Heights of students in a class {Doctors}

IRIS DATASET  $\rightarrow$  Petal length, Sepal Length

$\Downarrow$  Petal width Sepal width  
Researchers

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

Parameters:  $\mu \in \mathbb{R}$  = mean

$\sigma^2 \in \mathbb{R} > 0$  = Variance

$x \in \mathbb{R}$ .

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

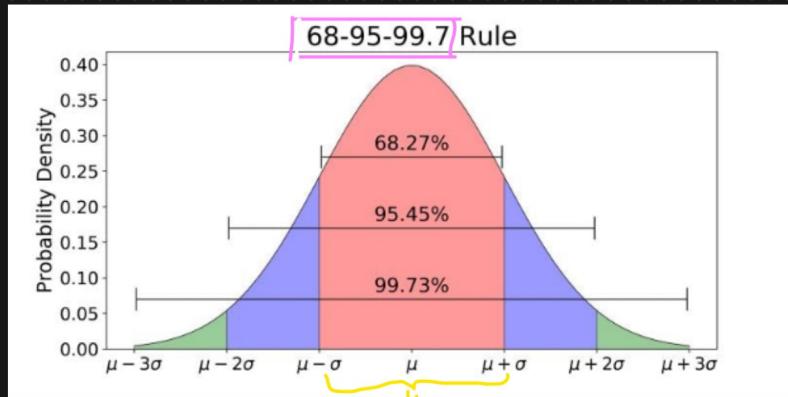
Mean of Normal / Gaussian

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

Variance

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

## Empirical Rule of Normal/Gaussian Distribution



QQ plot

$X = \{ \dots \} \Rightarrow$  Normal/Gaussian Distribution

• Probability

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

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

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

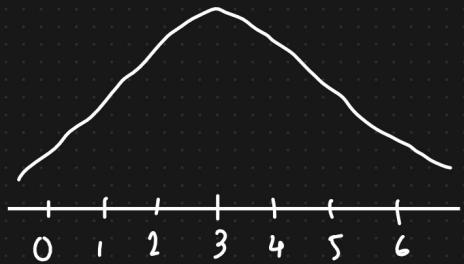
# ① Standard Normal Distribution

Z-score

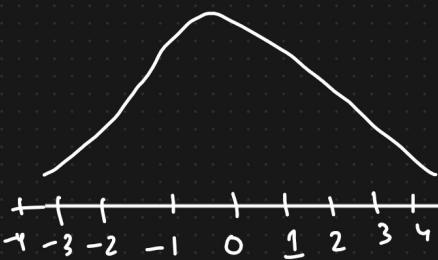
$$X = \{1, 2, 3, 4, 5\}$$

$$\mu = 3$$

$$\sigma = 1.414 \approx 1$$



$$\Rightarrow \begin{array}{l} \mu = 0 \\ \sigma = 1 \end{array}$$



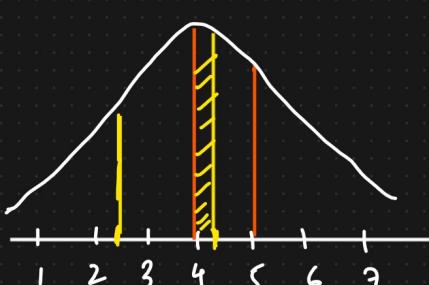
$$X = \{1, 2, 3, 4, 5\}$$

$$Z\text{-score} = \frac{x_i - \mu}{\sigma} \quad Y: \{-2, -1, 0, 1, 2\}$$

$$\textcircled{1} \quad \frac{1-3}{1} = -2 \quad \textcircled{3} \quad \frac{3-3}{1} = 0$$

$$\textcircled{2} \quad \frac{2-3}{1} = -1 \quad \textcircled{4} \quad \frac{4-3}{1} = 1$$

$$X \sim \text{SND}(\mu = 0, \sigma = 1)$$



$$\mu = 4$$

$$\sigma = 1$$

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

$$x_i = 4.25$$

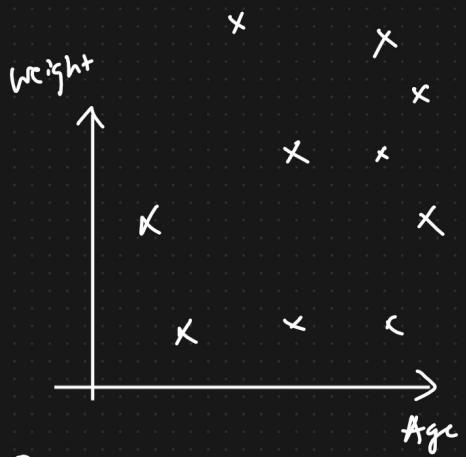
$$Z\text{-score} = \frac{4.25 - 4}{1} = 0.25$$

$$x_i = 2.5$$

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

Eg: Dataset

	Age	Years	kg	Cms	TINR
			Weight	Height	Salary
	24		70	175	40K
	25		60	160	50K
	26		55	180	60K
	27		40	130	30K
	30		30	175	20K
	31		25	180	70K
	32		↓	↓	↓



① Clustering Algorithms

② Linear Regression

③ Logistic Regression

Standardization  $\Rightarrow$  ML Models

$$Z\text{-score} = \frac{x_i - \text{Mean}}{\sigma}$$

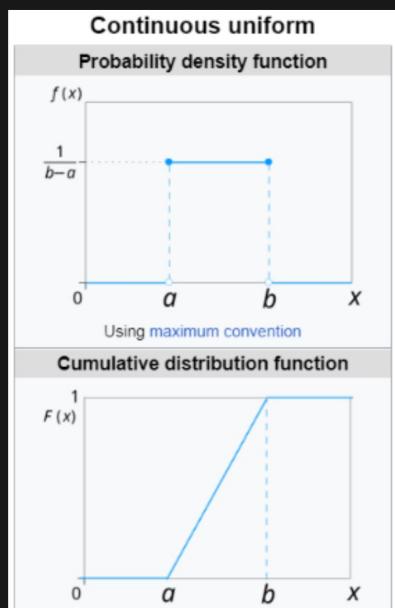
## Uniform Distribution

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

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

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

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

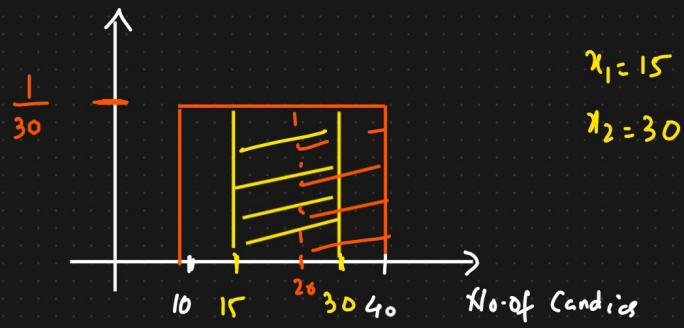
$$\text{Variance} = \frac{1}{12}(b-a)^2 //$$

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

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)



$$x_1 = 15$$

$$x_2 = 30$$

No. of Candies

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

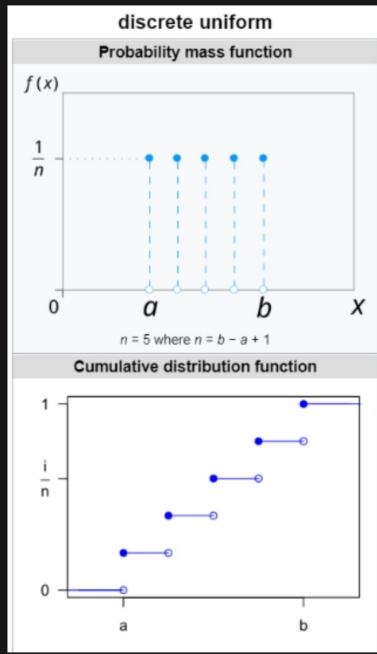
$$= (30 - 15) * \frac{1}{30}$$

$$= 0.5 \text{ //}$$

$$\Pr(X > 20) = (40 - 20) * \frac{1}{30} = 0.666 = 66\% .$$

## (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  $\pi$  Fair dice  $\{1, 2, 3, 4, 5, 6\}$

$$\frac{1}{n} \Rightarrow n = b - a + 1 = 6 - 1 + 1 = 6$$

$$Pr(1) = \frac{1}{6}$$

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

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

$$\vdots = \frac{1}{6}$$

$\frac{1}{6}$

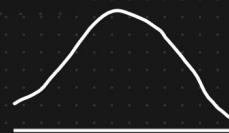
Notation  $U(a, b)$

Parameters  $a, b$  where  $b > a$

PMF  $\frac{1}{n}$

Mean  $\rightarrow \frac{a+b}{2}$   
Median  $\rightarrow \frac{a+b}{2}$ .

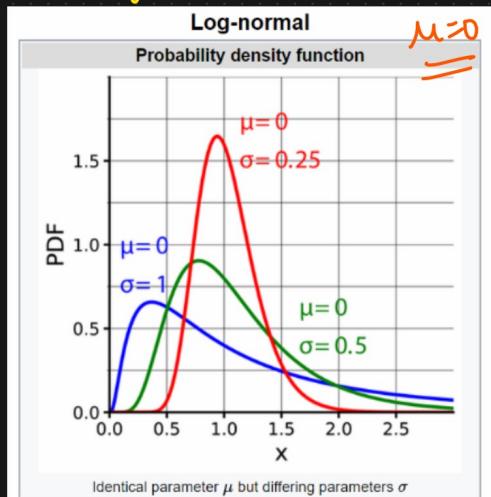
# ① Log Normal Distribution



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



$$X \sim \text{Log Normal Distribution}(\mu, \sigma)$$



$$Y \sim \ln(X) = \text{Normal Distribution}$$

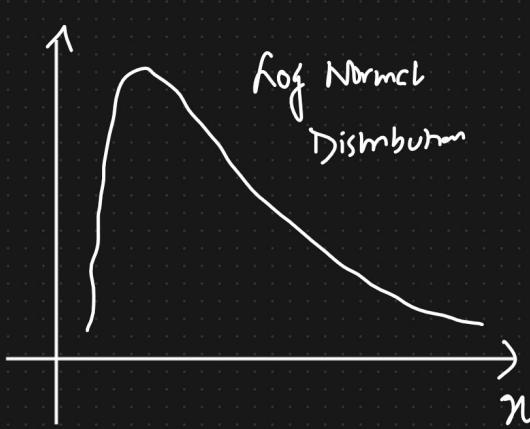
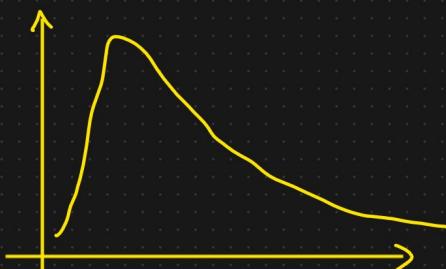


Natural log

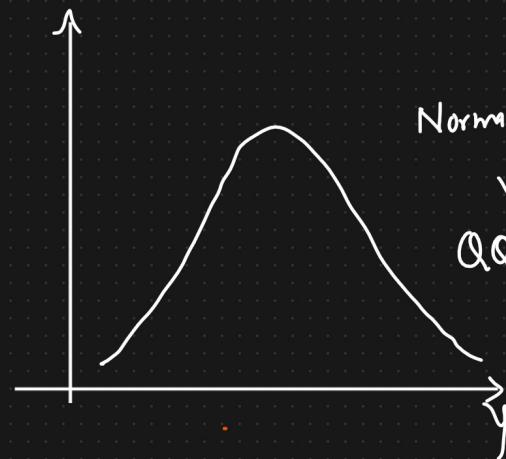
$$[\log_e]$$



$$X \sim \exp(Y) \Rightarrow \text{Log Normally Distributed}$$



$$\ln(x) \rightarrow$$



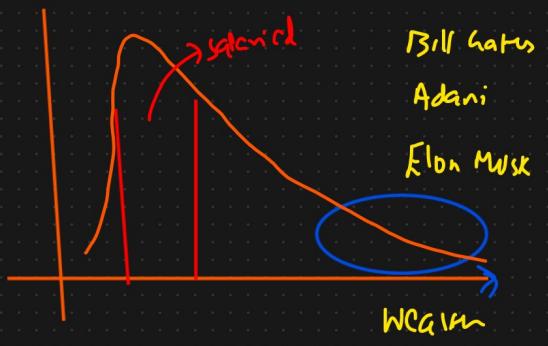
Normal Distribution



QQ plot

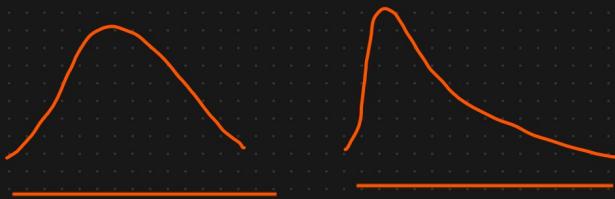
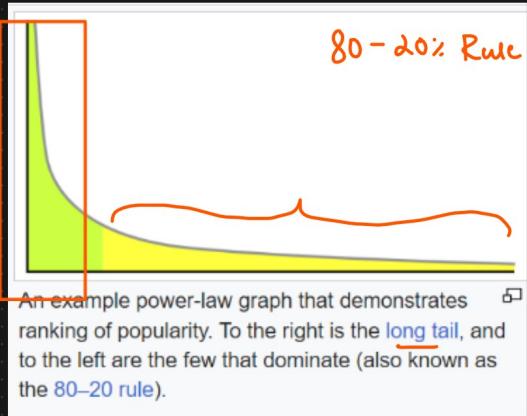
Eg:-

- 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

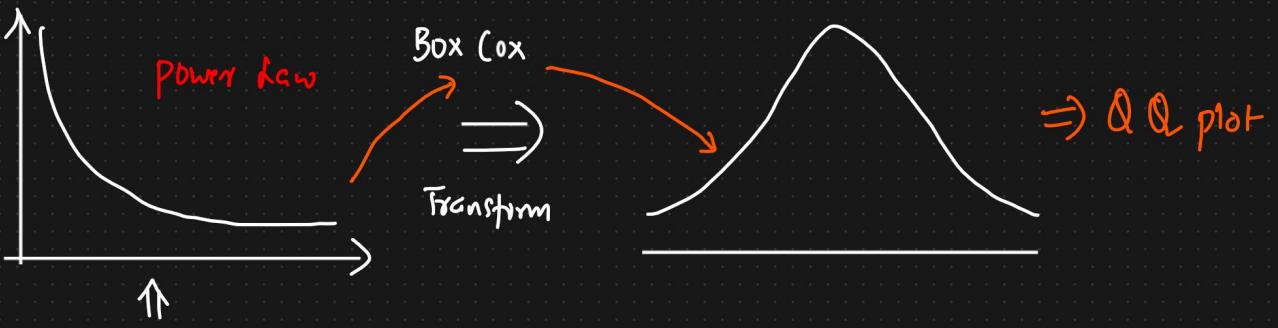
In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another



Eg: IPL

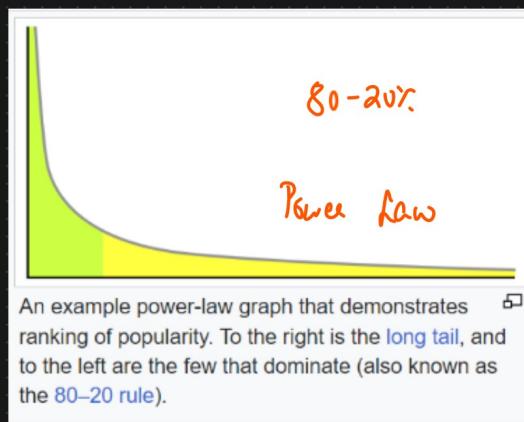
- ① 20% of Team is responsible for winning 80% match
- ② 80% of wealth are distributed with 20% of the total population
- ③ 80% of the total oil is with 20% of the nation
- ④ Frequencies of words in most languages.
- ⑤ 20% of the major defects fixes the 80% of upcoming defects in a slow product



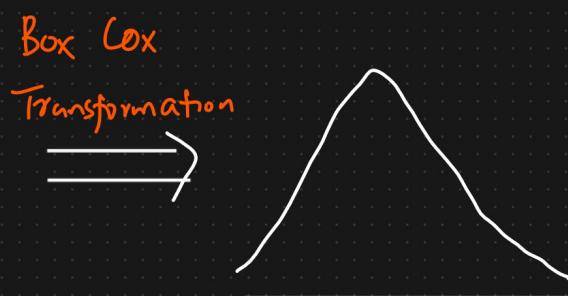
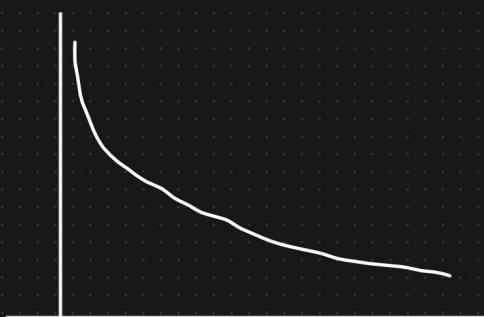
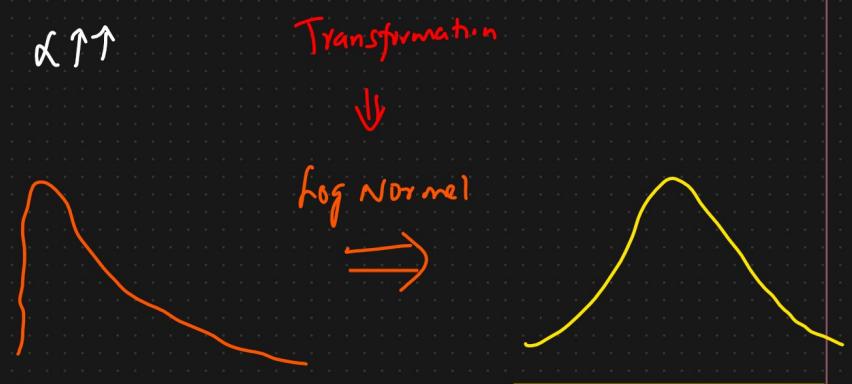
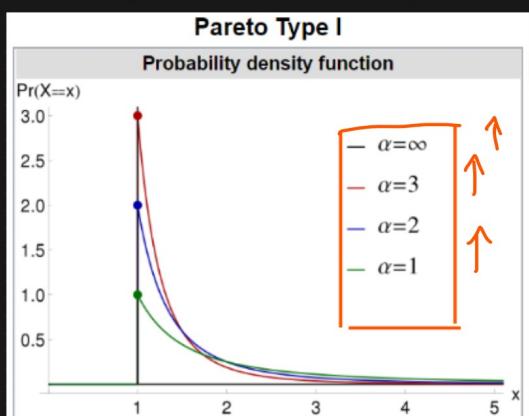


Pareto Distribution [80-20%]

# ① Pareto Distribution [Non Gaussian Distribution]



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 for. of defects can be solved if we  
Save 20% of defects

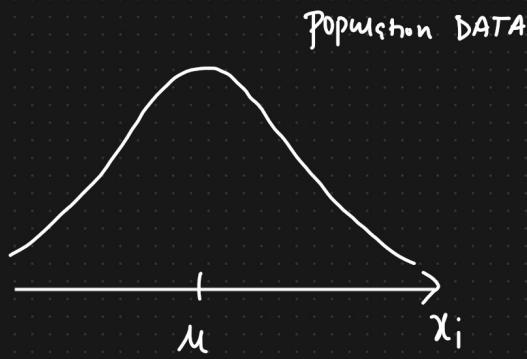
## ① Central Limit Theorem

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.

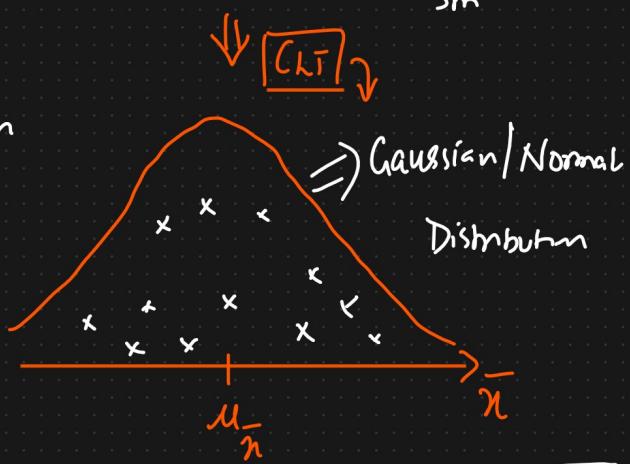
$n = \text{Sample Size} \Rightarrow$  any value

$$① X \sim N(\mu, \sigma)$$



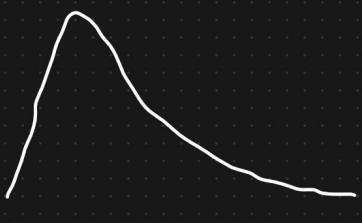
$$\begin{aligned} S_1 &= \{x_1, x_2, x_3, \dots, x_n\} = \bar{x}_1 \\ S_2 &= \{x_2, x_3, \dots, x_n\} = \bar{x}_2 \\ S_3 &= \dots \\ S_4 &= \dots \\ &\vdots \\ S_m &= \dots \end{aligned}$$

Sampling distribution  
of the mean



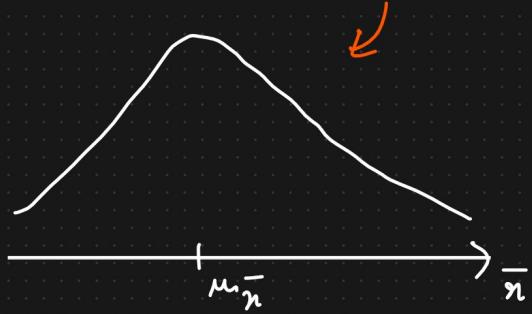
$\Rightarrow n > 30 \Rightarrow$  sample size

$$② X \not\sim N(\mu, \sigma)$$

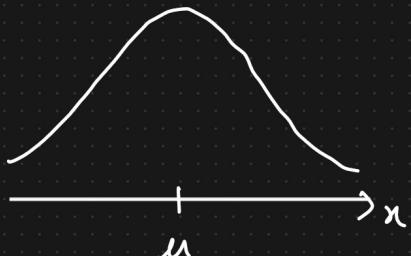


$\Downarrow$  CLT

$$\begin{aligned} S_1 &= \dots \\ S_2 &= \dots \\ &\vdots \\ S_m &= \dots \end{aligned}$$



## ① Normal Distribution

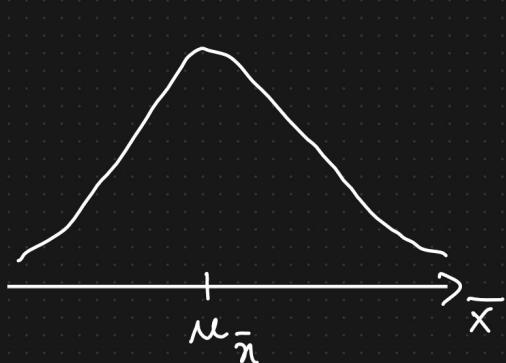


$$X \sim N(\mu, \sigma)$$



$\sigma$  = population std  
 $\mu$  = population mean  
 $n$ : sample size

## Sampling Distribution of mean



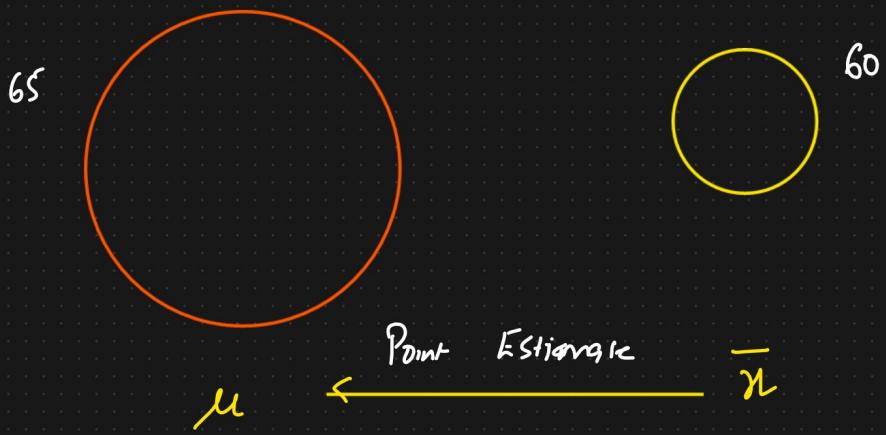
$$X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Estimate: If 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

$$[55 - 65] \Rightarrow \text{Sample Mean}$$

