What is Statistics

Statistics is a branch of Mathematics that involves Collecting, analysis, interpreting and presenting data. It provide tools and methods to understand and make sense of large amount of data and to draw conclusions and make decisions based on the data.

In practice, statistics is used in a wide range of field, such as bussiness, economics, social science, medicine and engeneering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions and make forecasts and predictions.

Example :

① Business → Data analysis (Identifying customer behaviour) and Demand forecasting.

② Medical → Identify efficancy of new mediciens (Clinical trials), Identifying risk factor for diseases.

③ Goverment and Politics → Conducting surveys, Poiling.

④ Environmental science → Climate research

Types of Statistics

① Descriptive statistics

↳ Descriptive statistics deals with the collection, organization, analysis, interpretation and presentation of data. It focuses on summarizing and describing the main features of set of data, without making inferences or predictions about the larger population.

② Inferential Statistics

↳ Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occuring, hypothesis testing to determine if a certain claim about a population is support by the data, and regression analysis to examine the relationship between variables.

Population → Population is the entire group of individual

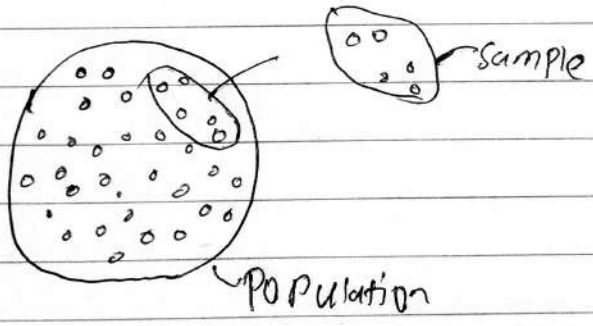
ex
① All People of India
② All the User of Netflix.

Sample → A small subset of Population that represents the whole.

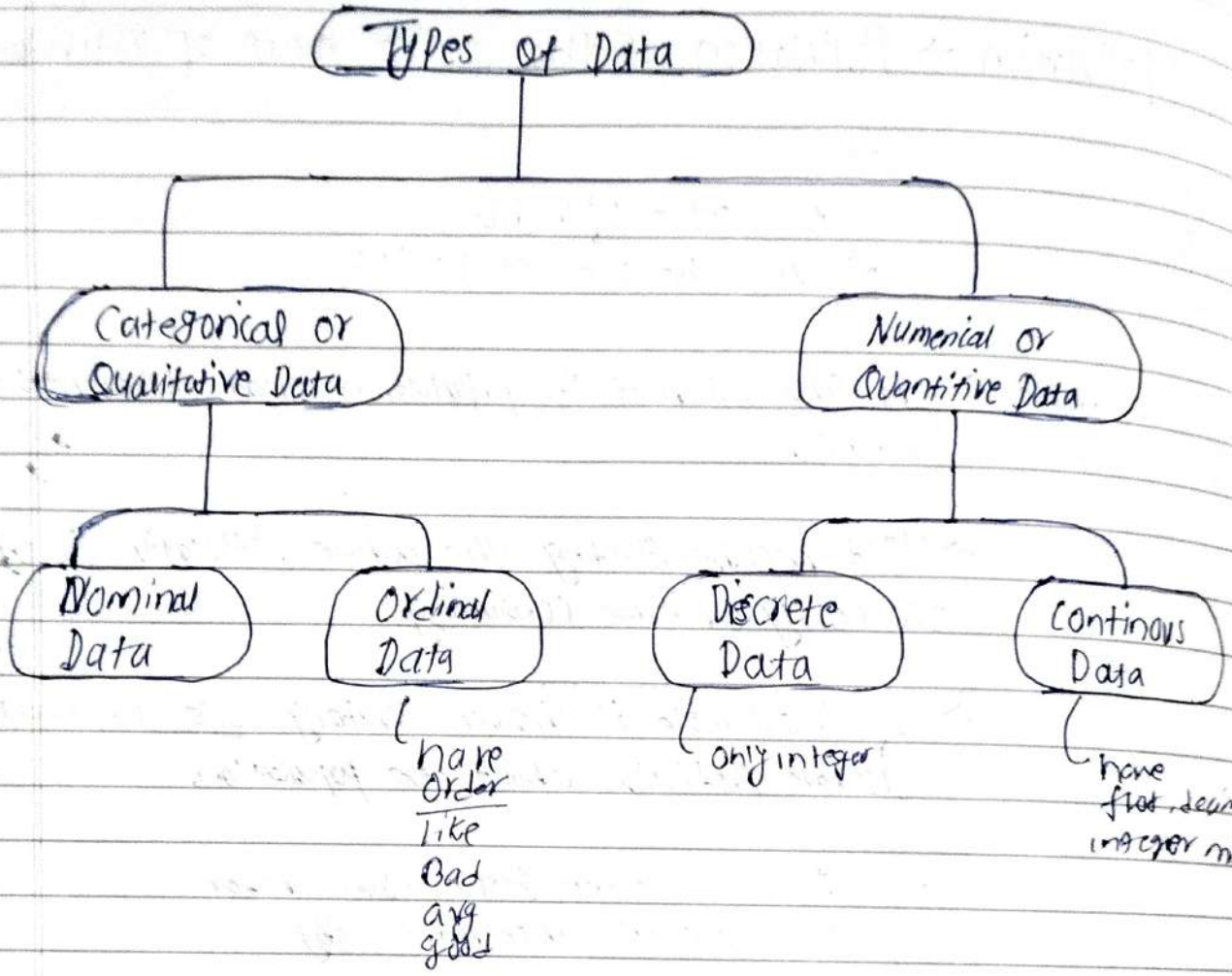
→ Used because studing the whole Population is often too costly or time consuming.

→ If the sample is choosen properly, we can make accurate prediction about the population.

ex ① 100 random People from bihar
② 80 Netflix User from fbg



Population → whole group
Sample → small part


Types of Data
Categorical or Qualitative Data
Numerical or Quantitive Data
Nominal Data
Ordinal Data
Discrete Data
Continous Data
have order like bad avg good
only integer
have float, decimal, integer no.

Measure Of central tendency

↳ A measure of central tendency is a stastical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.

(1) Mean

↳ The mean is the sum of all values in the dataset divided by the number of values.

Population mean	Sample mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = No of items in the Population	n = no of items in sample
Population mean is represented by μ (mew)	Sample mean is represented by x bar ($\bar{x}$)

(2) Median

↳ The median is the middle value in the dataset when the data is arranged in order.

③ Mode

↳ The mode is the value that appears most frequently in the Dataset

④ Weighted mean

↳ The weight mean is the sum of Products of each value and its weight, divide by the sum of the weights. It is used to calculate a mean when the data values in the dataset have different importantance or frequency.

⑤ Trimmed mean

↳ A trimmed mean is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called trimming percentage.

20000, 22000, 23000 | 27000, 28000, 30000, 32000, 35000, | 35000, 80000

↳ 36500

↳ full Icamean

25000, 28000, 30000 32000 35000

↳ 30000

↳ trimmed mean

Measure of Dispersion

↳ A measure of Dispersion is a statistical measure that describes the spread or variability of a Dataset. It provides information about how the Dataset is distributed around the central tendency (mean, median, mode) of the Dataset.

(1) Range

↳ The range is the difference between the maximum and minimum values in the Dataset. It simple measure of dispersion that is easy to calculate but can be affected by outliers.

Small range → Means all your values are close to each other - no much difference between min and max.

Interpretation
- low variation
- Data is stable and Predictable
- Indicate Consistency.

large range → Means values are far apart - big gap between min and max

Interpretation
- High variation
- Data is inconsistent and Unpredictable
- Indicates diversity or instability.

② Variance → The variance is the squared differences between each data point and the mean. It measures the average distance of each point from the mean is useful in comparing the dispersion of dataset with different means.

	x-mean	$(x-\text{mean})^2$
3	3-3	0
2	2-3	1
1	1-3	4
5	5-3	4
4	4-3	1

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$ Population variance

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$ Sample variance

Mean absolute Deviation

$$MAD = \frac{\sum |x_i - \bar{x}|}{n}$$

③ Standard Deviation → The standard deviation is the square root of the variance. It is a widely used measure of disperion that is useful in describing the shape distribution

$$\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$$ population SD

$$s = \sqrt{\frac{\Sigma (x-\bar{x})}{n-1}}$$ sample SD

④ Coefficient of variation

↳ Cofficient of variation (CV) is the ratio of the SD to the mean expressed as a ~~Percentile~~ percentage. It is used to compare the variability of data set with different means and is commonly used in field such as biology, chemistry and engineering.

→ CV is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is a dimensionless quantity that is expressed as percentage.

formula of CV

CV = Standard deviation / mean X 100%

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100$$

↓

CV yeh batata hai ke mean k respect me data kitne spread hai

Interpretation:

- It is expressed in Percentage
- Higher CV → more spread → more inconsistent data
- Lower CV → less spread → more consistent data

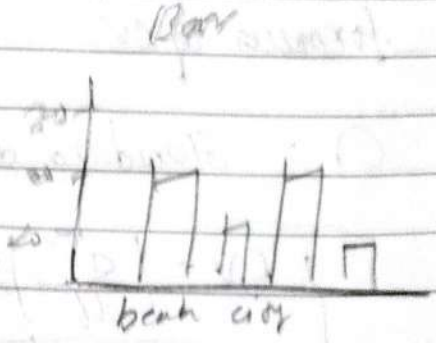
Graphs for univariate analysis

① Categorical – frequency Distribution table and Cumulative frequency.

A frequency distribution table is a table that summarizes the number of times [or frequency] that each values occur in dataset.

lets say we have a survey of 200 people & we ask them about their favourite types of vacation, which could be one of six categories: Beach, city, Adventure, Nature, Cruises or other.

Type of vacation	
Beach	40
City	60
Adventure	30
Nature	35
Cruise	20
Other	15



Quantiles

→ Quantiles are statistical measure used to divide a set of numerical data into equal sized group, with each group containing an equal number of observations.

Quantiles are important measure of variability and can be used to understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles used in statistical analysis including:

(a) Quartiles → Divide the Data into four equal parts, Q1 (25th percentile), Q2 (50th percentile) and Q3 (75 Percentile)

(b) Deciles → Divide the Data into 10 equal parts, D1 (10th percentile), D2 (20th Percentile)..... D9 (90th Percentile).

(c) Percentile → Divide the data into 100 equal parts, P1 (1st percentile) P2 (2nd Percentile)..... P99 (99th Percentile).

(d) quantiles → Divide the Data into [redacted] 5 equal parts

Things to remember while Calculating these measures:

1. Data should be Shorted from low to high
2. You are basically finding the location of an observation
3. They are not actual values in the Data
4. All other tiles can be easily derived from Percentiles

D13

Percentile → A Percentile is a statistical measure that represent the percentage of observation in a dataset that fall below a particular value.

ex 75th percentile is the value below which 75% of the observation in the dataset fall.

formula to calculate percentile value:

$$PL = \frac{P}{100}(N+1)$$

Where
PL = the desired value location
N = the total no of observation in the dataset
P = the percentile rank (expressed as a percentage)

Example
find the 75th percentile score for the below data.

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Step1 : sort the data (ASC)

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

$$PL = \frac{75}{100}(10+1) = \frac{3}{4} \times 11 = \frac{33}{4} = 8.25$$

96 - 98

96 + 0.25 (2) = 96.5

75th percentile = 96.5

Percentile of a value

$$\text{Percentile} = \frac{x + 0.5y}{n}$$

x = no of value below the given value

Y = No of value equal to the given value

N = total no of values in the dataset.

Or

78, 82, 84, (88), 91, 93, 94, 96, 98, 99

find percentile

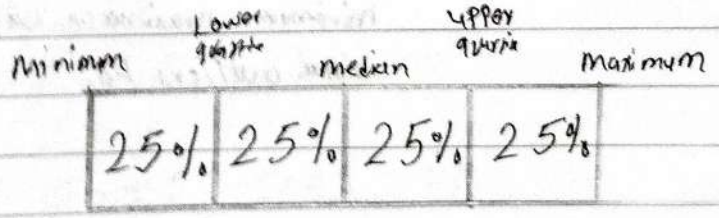
$$\text{Percentile} = \frac{3 + 0.5 \times 1}{10} = \frac{3.5}{10} = 0.35 \ (35\%)$$

5 Number Summary

The five numary summary is a descriptive statistic that provide a summary of a dataset. It consist of five values that divide the dataset into four equal parts, also know as quartiles. The five-number summary includes the following values:

1. Minimum values → The smalest value in the dataset
2. first Quartile (Q1) → The value that seprate the lowest 25% of the data from the rest of the data.
3. median (Q2) → The value that seprates the lowest 50% from the highest 50% of the data.
4. Third quarile (Q3) → The value that seprate the lowest 75% of the Data from the highest 25% of the Data.
5. maximum value → The largest value in the dataset.

The five number summary is often represented visualy using a box Plot which display the range of the Dataset, the median and the Quartiles. The five number summary is a Useful way to quickly summerize the center tendency, variability and distribution.

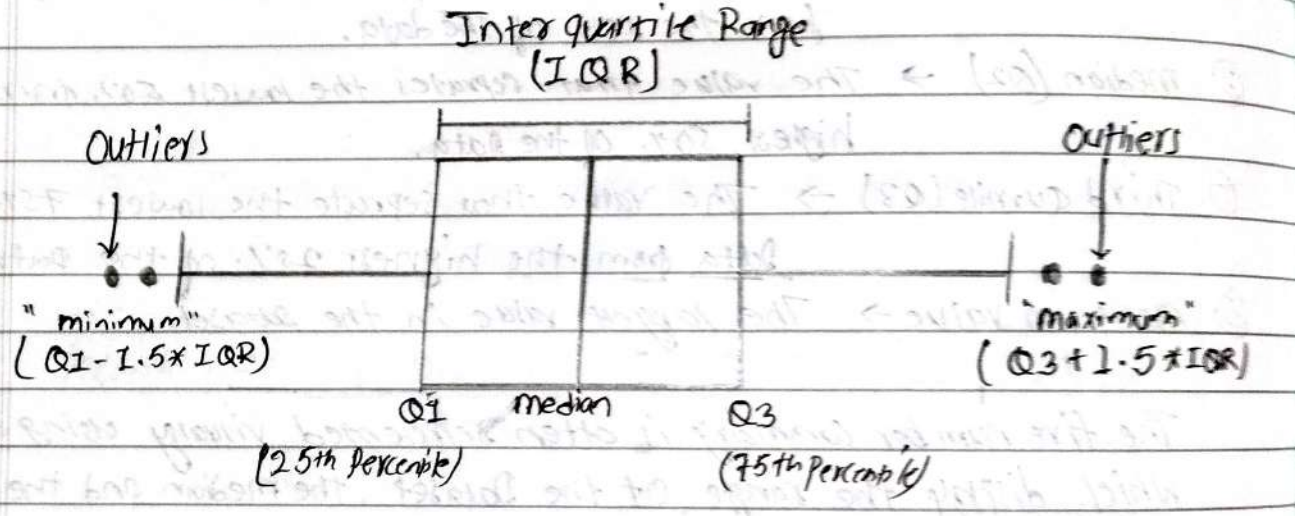


Inter quartile range (IQR)

↳ IQR is a measure of variability that is based on five number summary of a dataset. Specifically, the IQR is defined as the difference between third quartile (Q3) and the first quartile (Q1) of a dataset.

What is a box plot

→ A box plot, also know as box and whisker plot is a graphical representation of a dataset that show the Distribution of the data. The box plot displays a summary of the data, including the minimum and maximum value, the first quartile (Q1), the median (Q2), and the third quartile (Q3).

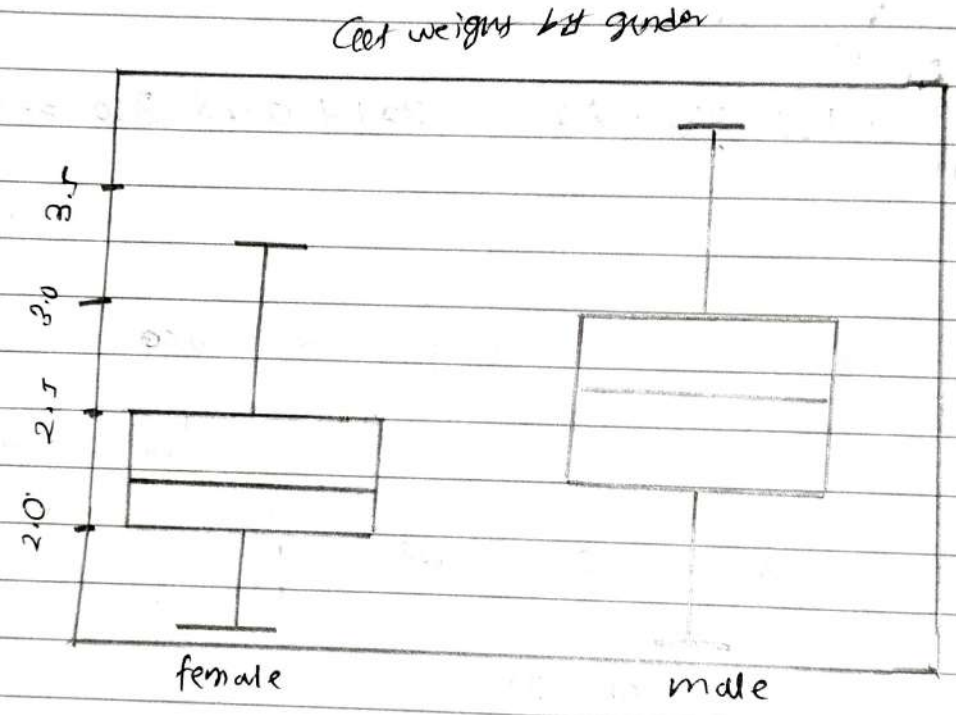


Minimum - Maximum se bahar wala sab value outliers ha;

Benifits of a box Plot

- ↳ easy way to see distribution of data
- ↳ Tells about skewness of data
- ↳ Can identify outliers
- ↳ Compare 2 category of data

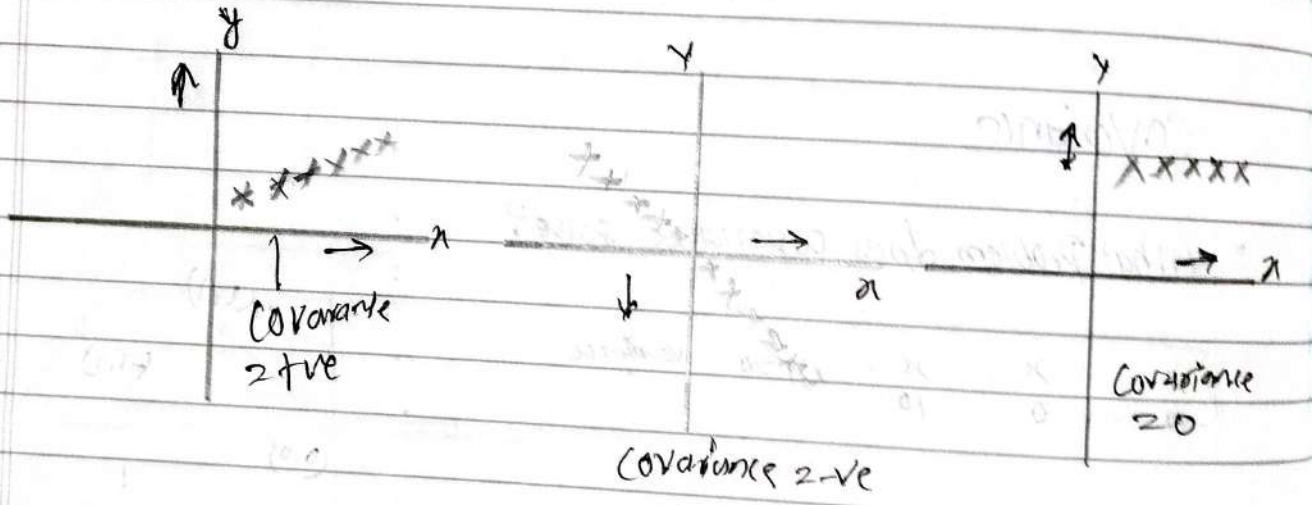
② Side by side box Plot.



- What is Covariance and How it is interpreted?

→ Covariance is a Stastical measure that Describe the Degree to which two variable are linearly related. It measures how much two variable change togeather, such that when one variable increase, does the other variable also increase, or does it decrease?

If the Covariance between two variable is Positive, it means the variable tend to move togeather in the same direction. If Covariance is negative, it means the variable tend to move in the opposite directions. A covariance of zero indicates that the variable are not linearly related.



How is it calculated?

Covariance formula

Population	Sample
$\sigma_{xy} = \frac{\sum (x-\mu_x)(y-\mu_y)}{N}$	$S_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1}$
$x, y \rightarrow$ The value of x & y in the Population	$x, y \rightarrow$ The value of x & y in the sample
$\mu_x, \mu_y \rightarrow$ The Population mean of x and y.	$\bar{x}, \bar{y} \rightarrow$ The sample mean of x and y.
$N \rightarrow$ Total no of Observation	$n \rightarrow$ Total number of Observation.

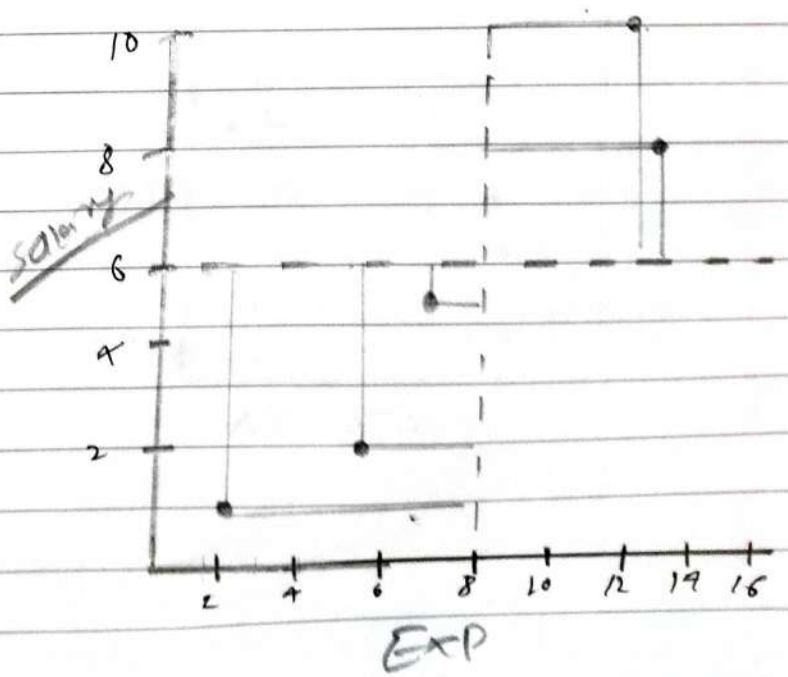
Exp (x)	Salary (y)	x-x mean	y-y mean	(x-x mean)x(y-y mean)
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	-1
12	12	4	6	24
13	10	5	-4	20

$\bar{x} = 8 \quad \bar{y} = 6$

$$\sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

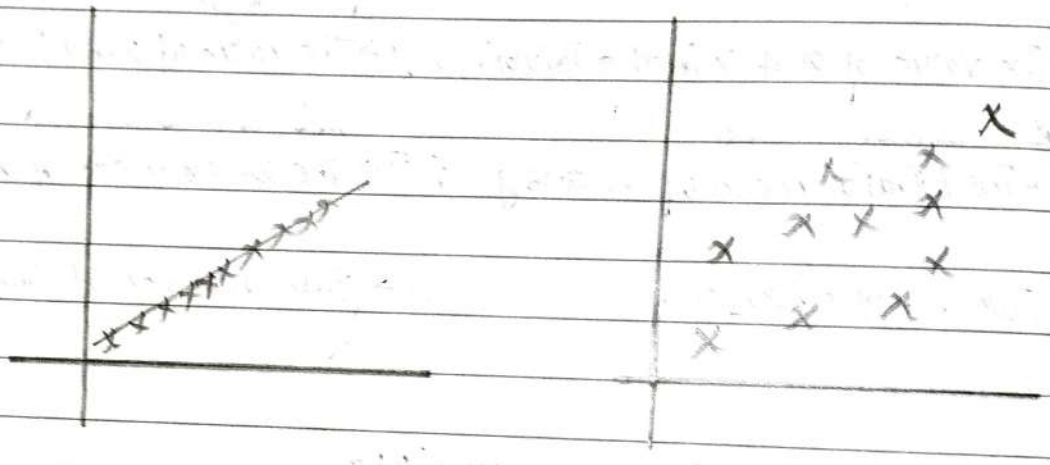
$$\frac{85}{4}$$

$$\text{COV} = 21.5$$



Disadvantage of using Covariance

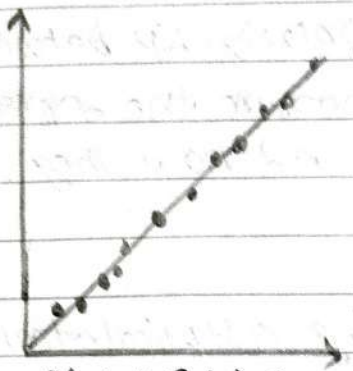
One limation of Covariance is that it does not tell us about the strength of relationship between two variables, since the magnitude of covariance is affected by the scale of the variable.



Note → Dono true hai but ek ka relationship bohat accha hai aur ek sirf true hai.

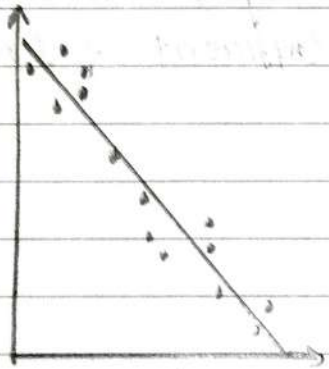
Correlation

① What Problem does the Correlation solve?

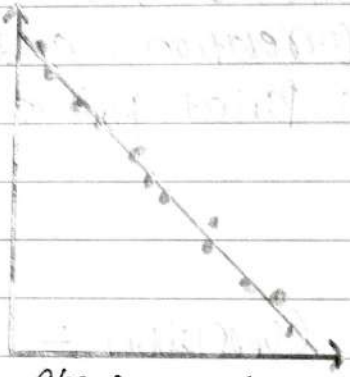


Strong Positive Correlation

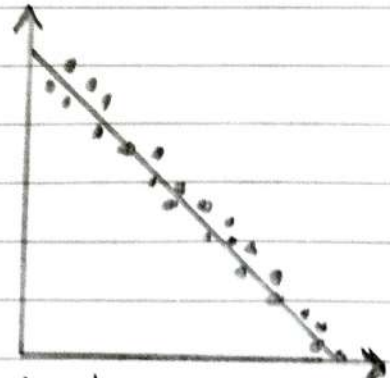
Weak Positive Correlation



Weak negative Correlation



Strong negative Correlation



Modrate Negative Correlation

No correlation

② What is Correlation?

→ Correlation refers to a statistical relationship between two or more variable. Specifically, it measure the degree to which variable are related and how they tend to change together

Correlation is often measured using a stastical tool called the Correlation Coefficient, which range from -1 to 1. A correlation cofficient of -1 indicates a perfect negative correlation, a correlation coefficient of 0 indicates no correlation, and a correlation cofficient of 1 indicates a perfect positive correlation.

$$\text{Correlation} = \frac{\text{Cov}(x, y)}{\sigma x * \sigma y}$$

+1 = Perfect Positive corr
0 = No correlation
-1 = Perfect negative correlation

Correlation and Causation

The phrase "correlation does not imply causation" mean that just because two variable are associated with each other, it does not necessarily mean that one causes the other. In other words, a correlation between two variables does not necessarily imply that one variable is the reason for the other variables behaviour.

Suppose there is a positive correlation between the number of firefighters present at a fire and the amount of damage caused by the the fire. On might be tempted to conclude that the presence of firefighters causes more damage. However, this correlation could be explained by a third variable - the serenity of the fire. More servere fire might require more firefighters to be present, and also cause more damage.

Thus, while correlation can provide valuable insights into how different variable are related, they cannot be used to establish causality. Establishing causality often requires additional evidence such as experiments, randomized controlled trials, or well-designed observational studies.

Random Variable

- What is Algebric variable

In algebra a variable is like x, is an unknown value.

$x + 5 = 10$

$x = 5$

- What is Random variables in Statistic and Probability?

↳ A random variable is set of possible values from a random experiment.

Coin toss → H, T

$X = \{1, 0\}$

H = 1, T = 0

Dice → 1 – 4, 2 – 5, 3 – 6

$Y = \{1, 2, 3, 4, 5, 6\}$

↳ Sample space

(1) What are Probability Distribution.

→ A Probability Distribution is a list of all possible outcome of random variable along with their corresponding probability values.

Coin toss	H	T
Probability	$\frac{1}{2}$	$\frac{1}{2}$

Dice

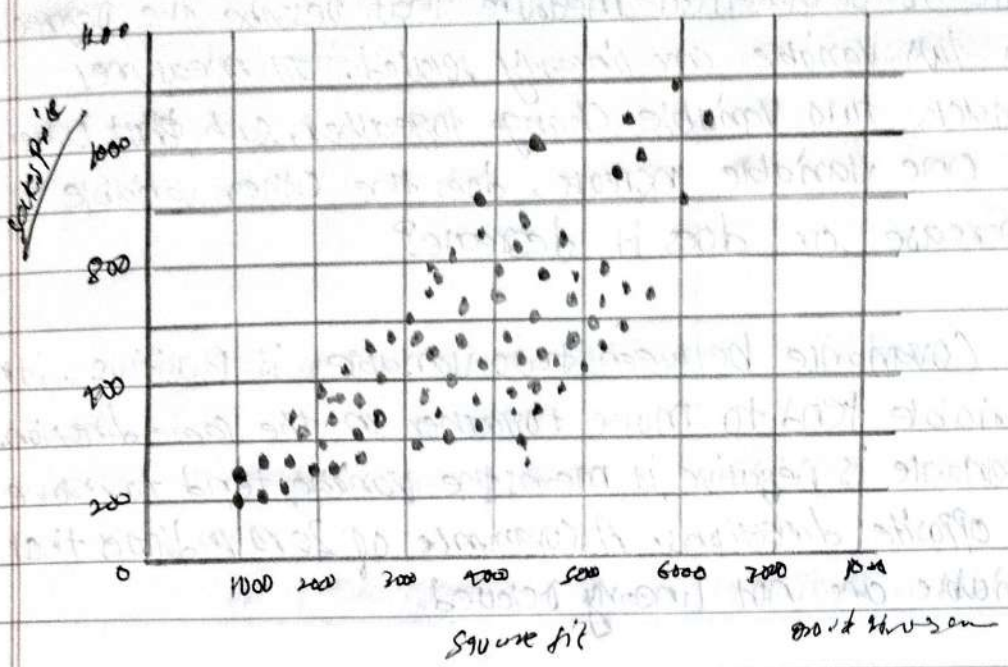
1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

2 dice

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

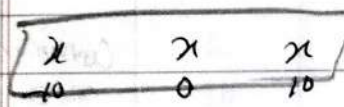
$2 \rightarrow 1/36$
$3 \rightarrow 2/36$
$4 \rightarrow 3/36$
$5 \rightarrow 4/36$
$6 \rightarrow 5/36$
$7 \rightarrow 6/36$
$8 \rightarrow 5/36$
$9 \rightarrow 4/36$
$10 \rightarrow 3/36$
$11 \rightarrow 2/36$
$12 \rightarrow 1/36$

Scatter Plot

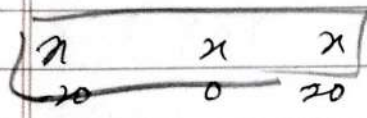


Covariance

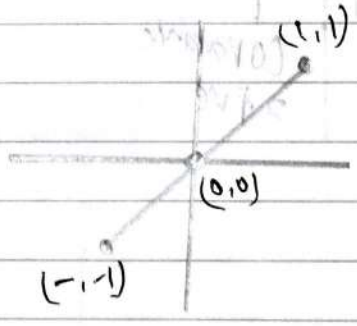
- What Problem does covariance solve?



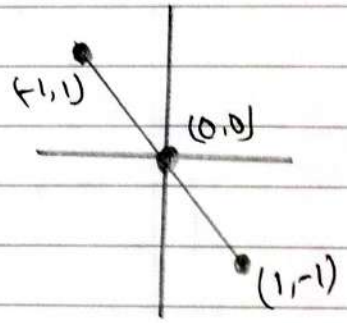
$m = 0$ variance



$m = 0$



$\frac{1^2 + 0^2 + 1^2}{2} = \frac{2}{2}$



$\frac{1^2 + 0^2 + 1^2}{2} = \frac{2}{3}$

Note → Covariance only tells the direction of relationship (Positive or Negative) But not the strenght, because it depends on the scale of variable.

Problem with Distribution

→ In many scenarios, the number of outcomes can be much larger and Hence a table would be tedious to write down. Worse still, the number of possible outcomes could be infinite, in which case, good luck writing a table for that.

Example → Height of people, Rolling 10 dice together.

Solution → function ?

→ What if we use a mathematical function to model the relationship between outcome and probability?

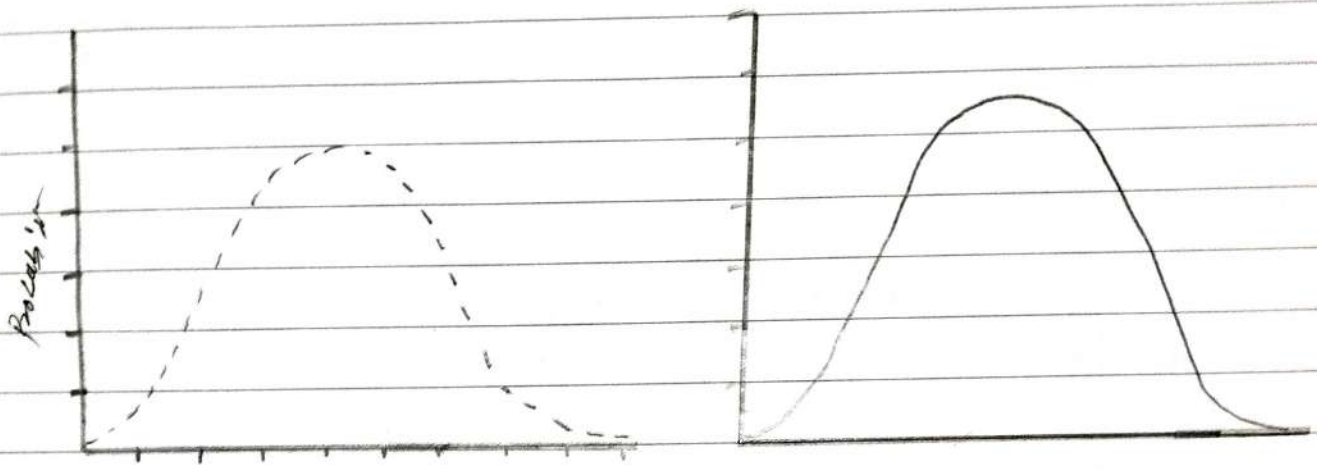
x → outcome
y → Probability

$y = f(x)$

NOTE → A lot of time Probability Distribution and Probability Distribution functions are used interchangeably.

Types of Probability Distribution (PDf)

- Descrete random variable
- Continuous Random Variable



Why are Probability Distribution important ?

↳ Gives an Idea about the shape / distribution of the Data.
↳ And if our data follows a famous Distribution we automatically know alot about the data.

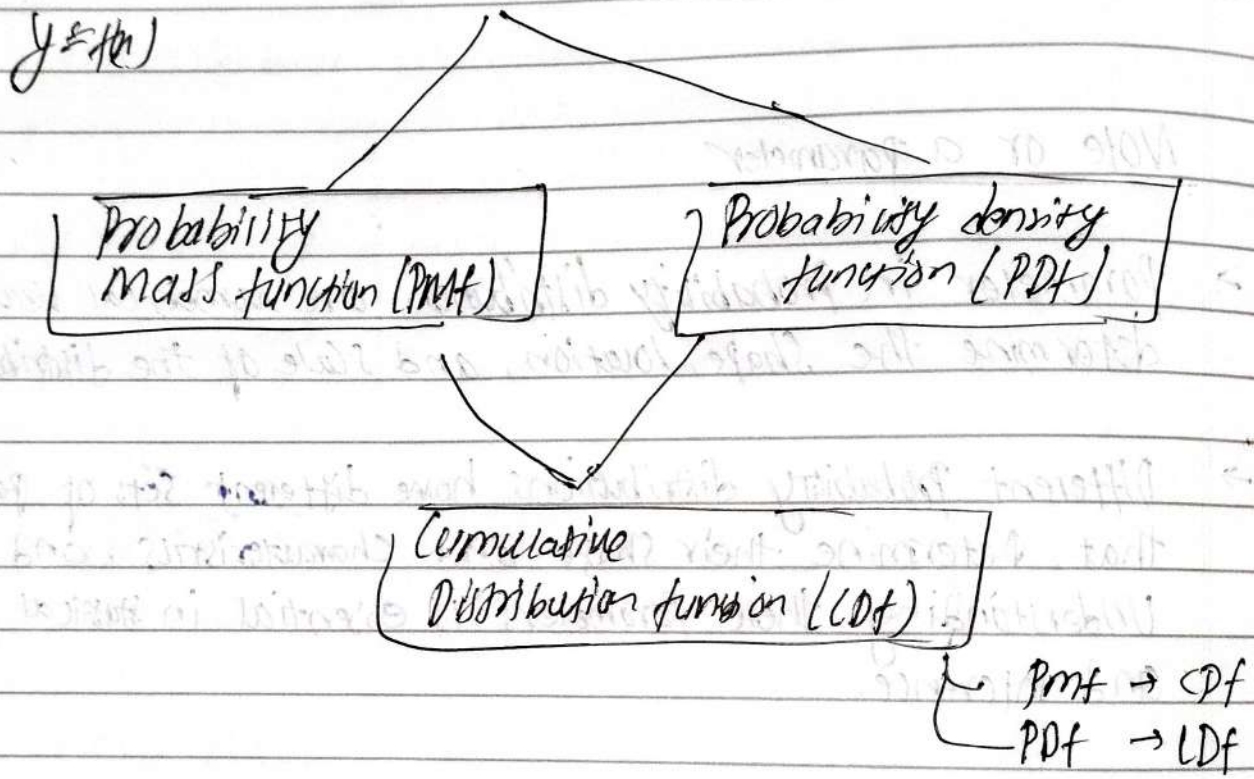
Note or a Parameter

→ Parameter in Probability distribution are numerical values that determine the shape, location, and scale of the distribution.

→ Different Probability distributions have different sets of Parameter that Determine their shape and Characteristics, and Understanding these Parameters is essential in statistical analysis and inference.

Probability Distribution function

A Probability distribution function is a mathematical function that describes the probability of obtaining different values of random variable in a particular probability distribution.



Probability Mass function (PMF)

PMF stands for Probability mass function. It is a mathematical that describe the Probability distribution of a discrete random variable.

The PMF of discrete random variable assigns a probability to each possible value of the random variable. The Probabilities assigned by the PMF must satisfy two conditions:

(a) The probability assigned to each value must be non-negative (ie., greater than or equal to zero).

(b) The sum of the probabilities assigned to all possible value must equal 1.

$y = f(x)$ $\quad y = \begin{cases} \frac{1}{6} & \text{if } x \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise} \end{cases}$

PMF $\rightarrow$ $y = \begin{cases} 1/36 & x \in \{2, 12\} \\ 2/36 & x \in \{3, 11\} \\ 0 & \text{otherwise} \end{cases}$

Note $\Rightarrow$ PMF se ek descrete random variable ke particular value ke probability nikal sakte hain.

Cumulative Distribution function (CDf) of Pmf

→ The Cumulative distribution function (CDf) $f(x)$ describes the Probability that a random variable X with a given Probability distribution will be found at avalue less than equal to x.

$$\left[f(x) = P(x \leq x) \right]$$

Discrete / Continuous

Note ⇒ CDf batata hai ke random variable ke kisi value se chotti ya uske barabar value aane ki total Probability kya hai

Example

Agar ek fair dice hai → x = number on dice.

Possible values : 1, 2, 3, 4, 5, 6

Pmf : $P(X = x) = \frac{1}{6}$

Toh CDf hoga :

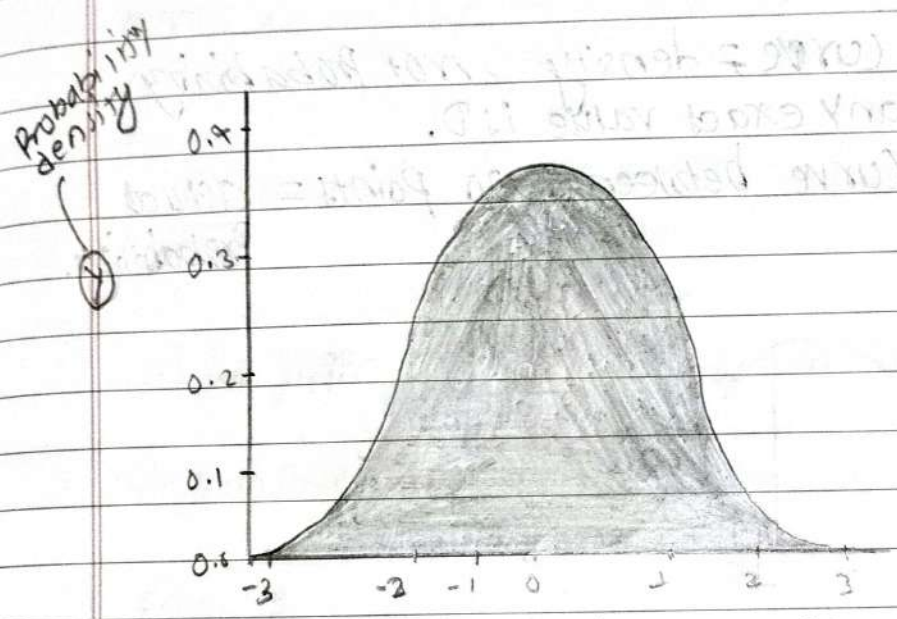
- $f(1) = P(X \leq 1) = \frac{1}{6}$
- $f(2) = P(X \leq 2) = \frac{2}{6}$
- $f(3) = P(X \leq 3) = 3/6$
- $f(6) = P(X \leq 6) = 1$

Matlab

→ CDf batata hai ke dice ka number 3 ya usse chotta value aane ka Chance = 3/6 = 0.5 (50%)

Probability Density function

→ PDf Stands for probability Density function. It is a mathematical function that describe the probability distribution of a continous random variable.



Why "Probability Density" not just "Probability"?

↳ ~~Probability~~ Ke jagaa probability Density use karte hai kyuki continous ma ek exact value ka probability 0 hota hai. Jaise agar hame ek particular value like 2.5 ka probability chaiyea, aur 1 aur 2 ke beech infinite values hai to.

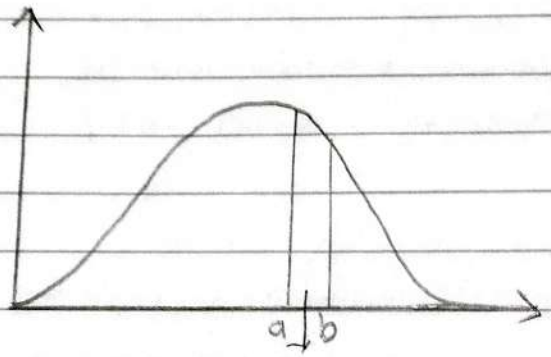
$$\frac{1}{\infty} \approx 0$$

Isleya ham probability nahi, probability density use karte hai.

- What is Probability Density?

Probability Density ⇒ ek function (curve) jo batata hai kisi value ke ass-pass Probability kitne "packed" hain

- Height of the curve = density, not Probability
- Probability of any exact value is 0.
- Area under curve between two points = actual Probability.



$P(a \leq X \leq b)$

In another words

→ Probability Density. Describe how dense the Probability is at a certain value of a random variable, but not the actual probability at a single point. for continous variable, the probability at any exact value is zero; instead you calculate the probability for a range by finding the area under the curve between two points.

Note → Probability Density batata hai ke 2 value ke beech (Point) na koi ek value aane ka probability kya hai.

- How to calculate Probability then?

Ham koi ek exact point ka probability nikalenge to 0 hoga isleya ham PDF nikalte hai jo bata a hai ke 2 points ke beech ma aana ka ek no ka probability kya

PDF formula: $P(a < X \leq b) = \int_a^b f(x)dx$

- f(x) ke jagha uska formula laga do (like distribution hai, uniform normal / exponential etc).
- Integral solve karo → jo number aayega, (wahi probability hai)

Example

Uniform [1,5], range 2-3:

$f(x) = 0.25 \Rightarrow P(2 \leq X \leq 3) = \int_2^3 0.25\,dx = 0.25 \times (3-2)$

$= 0.25$

Yahi hai probability $\Rightarrow 0.25$ ..

- How graph is calculated?

↳ Raw Data hai.
↳ Pata nahi kaunsa distribution follow karta hai
↳ Density estimation use karte hai taaki probability Distribution ka shape samajh aaye

[redacted]

Density estimation

Density estimation ka matlab hai data se probability Distribution ka estimate banana. Matlab, thumare pass sirf Data Points hain, aur tumhe nhi pata ke Data konsa theoretical distribution follow karta hai (like normal, uniform, exponential etc). toh ham ek curve banate ho jo bataye ki data probability kaha zada hai aur kaha kam.

Density estimation → raw data se distribution ka shape nikalne ka tarika.

- Density estimation can be used for a variety of purpose such as hypothesis testing, data analysis, and data visualization. It is particular useful in area such as machine learning, where it is often used to estimate the probability distribution of input data or model the likelihood of certain events or outcomes.

There are various method for density estimation, including parametric and non-parametric approaches.

Commonly used techniques for density estimation include kernal density estimation (KDE), histogram estimation, gaussian mixture model (GMMs) The choice of method depends on the specific characteristics of the data and the intended use of the density estimate.

Parametric Density Estimation

→ Parametric density estimation is a method of estimating the probability density function (pdf) of a random variable by assuming that the underlying distribution belongs to a specific parametric family of probability distributions, such as normal, exponential or poission distribution.

Explaination:

→ Parametric Density Estimation steps:

- Data ko observe karen aur assume karen ki yeh kisi parametric distribution (eg. normal Distribution) ko follow karta hai → Plot histogram to assume distribution.
- Sample data se parameters calculate karen, for example mean (μ) aur standard deviation (σ) agar normal Distribution assume kar rahe ho
- Uske baad app us parametric function ko use karke data ke leyen probability density values nikal sakte hai

Jaise Normal Distribution ke liya $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)}{2\sigma^2}}$ har value k leyen

- In values ko plot karen to visualize the pdf.

NON-Parametric Density Estimation

But sometimes the distribution is not clear, it's not one of the famous distributions.

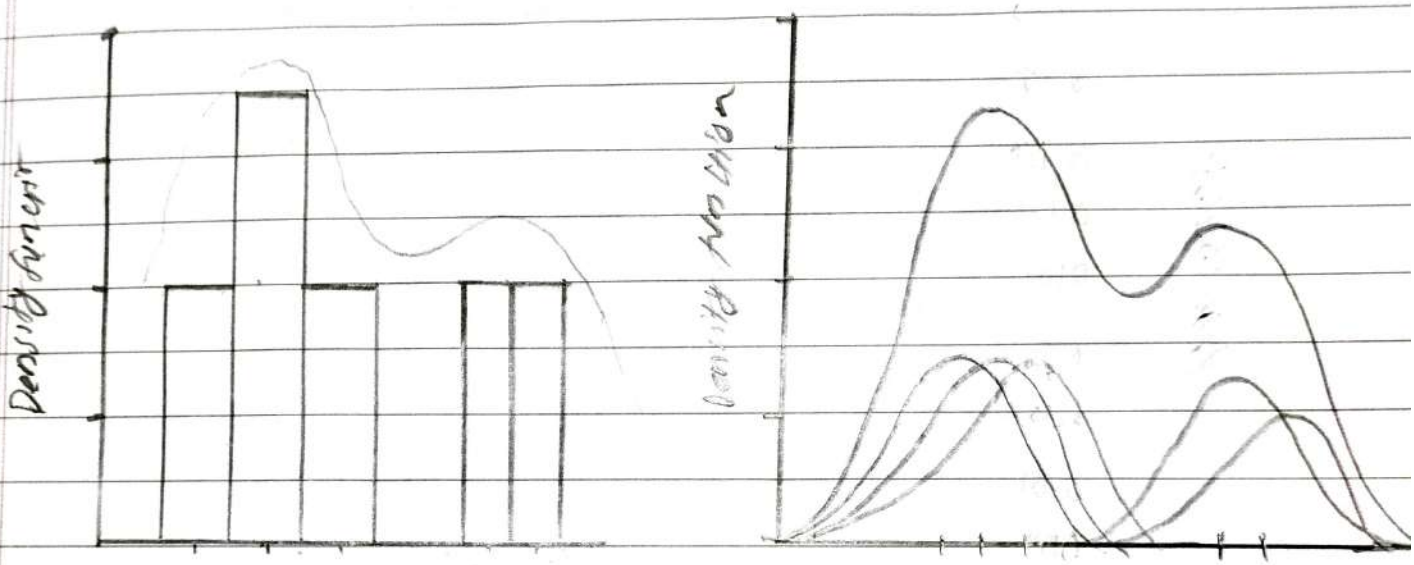
Non-Parametric density estimation is a statistical technique used to estimate the PDf of a random variable without making any assumption about the underlying distribution. It is also referred to as non-parametric density estimation because it does not require the use of predefined probability distribution function, as opposed to parametric method such as the gaussian distribution.

The NON-Parametric density estimation techniques involves constructing an estimate of the pdf using the available data. This is tipically done by creating a kernel density estimate.

Non Parametric density estimation has serval advantage over Parametric density estimation. One of the main advantage, which allows for more flexiable and accurate estimation in situations where the underlying distribution is unknown or complex. However, non-parametric density estimates compared to parametric methods.

Kernal Density Estimate (KDE)

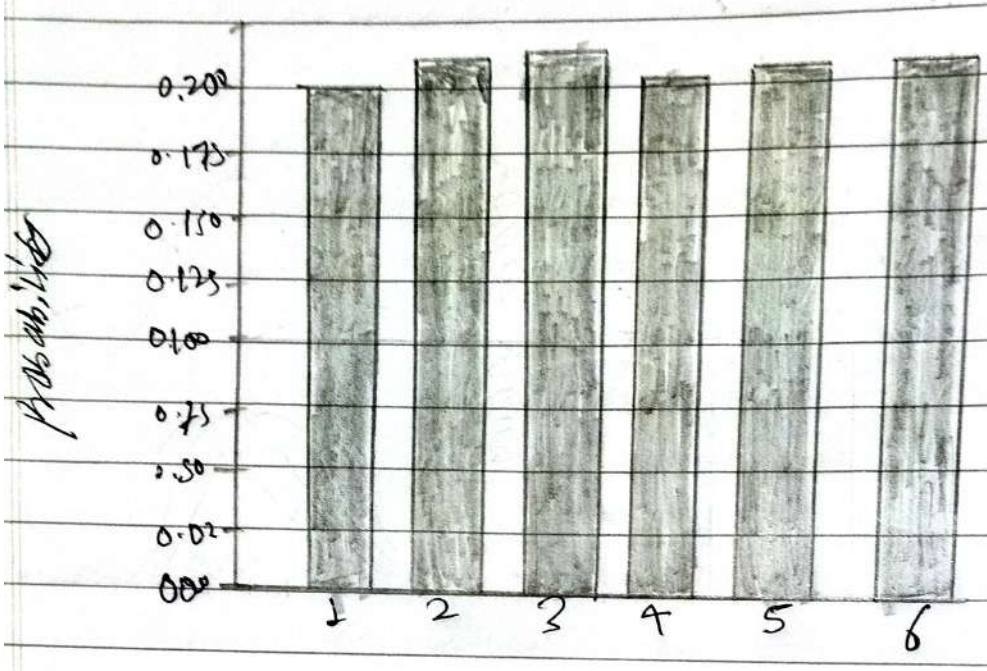
↳ The KDE technique involves using a kernal function to smooth out the data and create a continous estimate of the underlying density function.



Bandwith jithna jada hoga utha smooth hoga

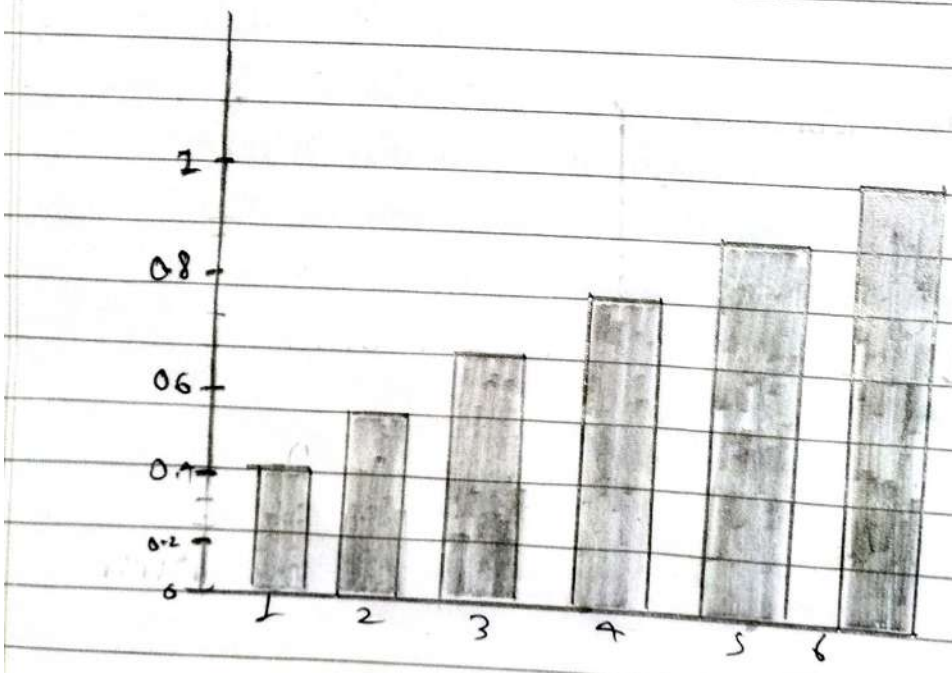
Cumulative Distribution function function (CDf of PDf)

Rolling a Dice and calculate PDf and CDf.



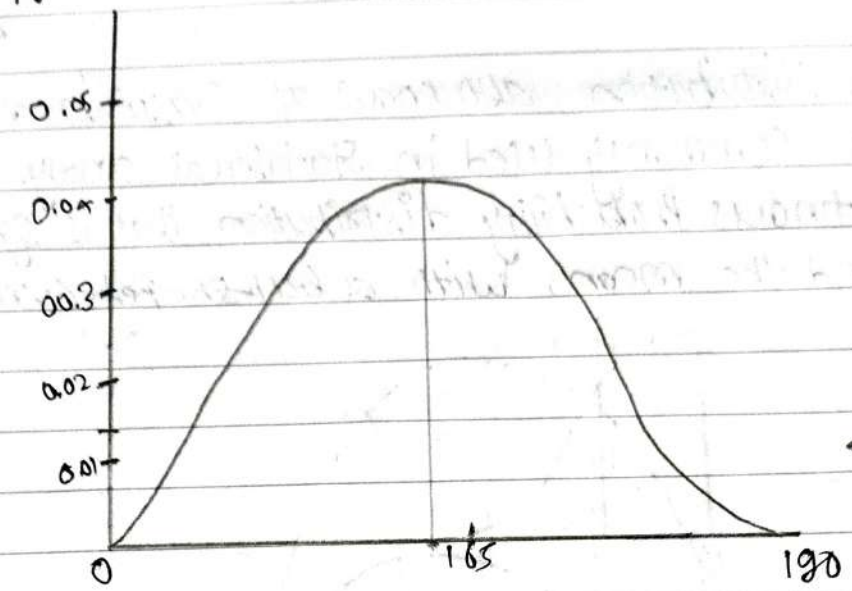
Pmf of dice

discrete values



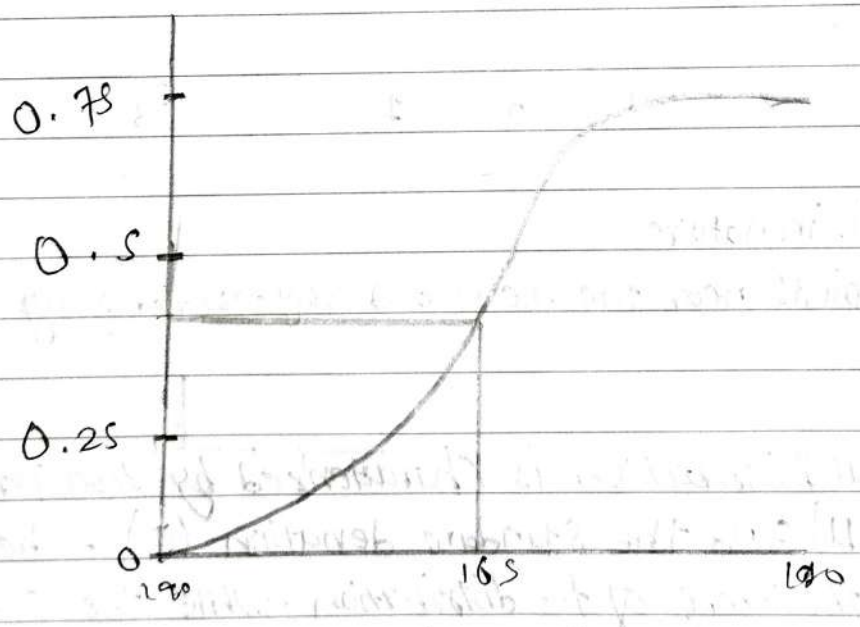
CDf of Dice and discrete random variable.

Probability Density



165 tone ka prob approx 40% hai

Cumulative Probability

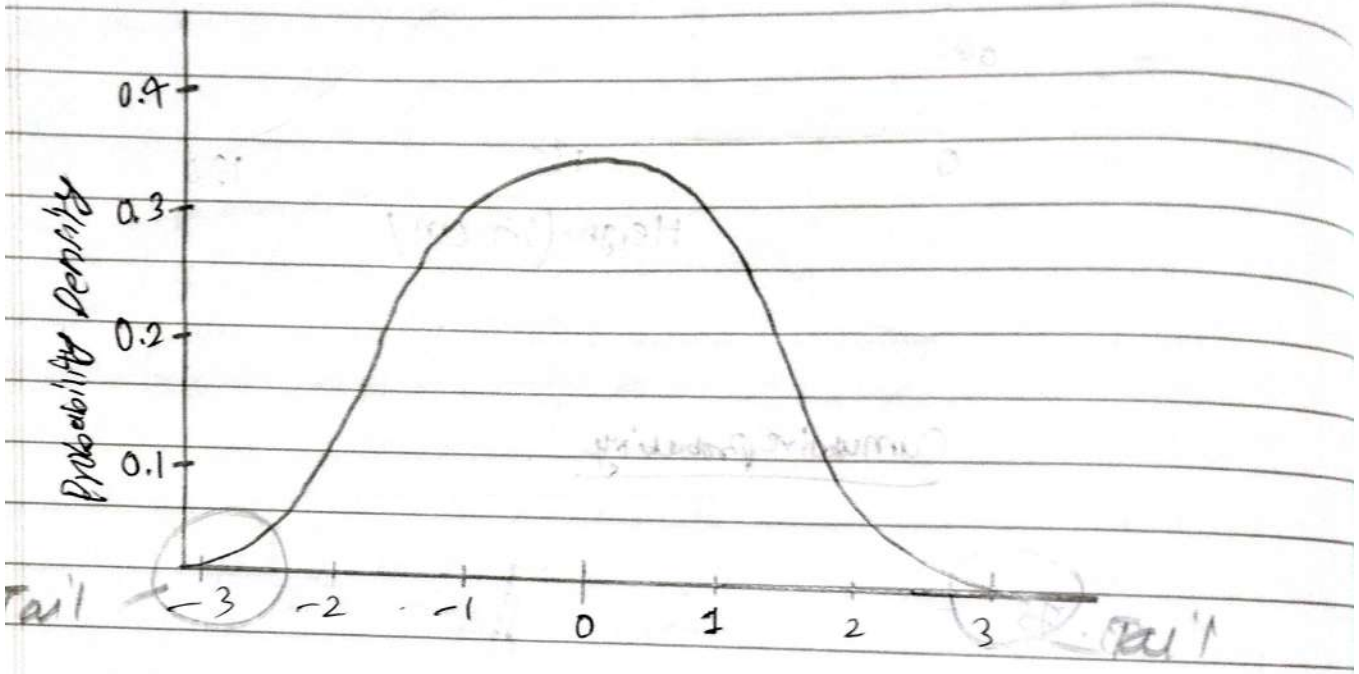


165 se kam honu ka chance 95% hai.

Normal Distribution.

bell curve

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



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

The Normal Distribution is Charaterized 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.

Denotes as:

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

Normal Distribution

classmate
Date
Page

Why it is so important?

↳ Commonly in Nature : Many Natural Phenomena follow a normal Distribution, such as the height of people, the weights of objects, the IQ scores of population and many more. Thus, the normal Distribution Provides a convenient way to model and analyse such data.

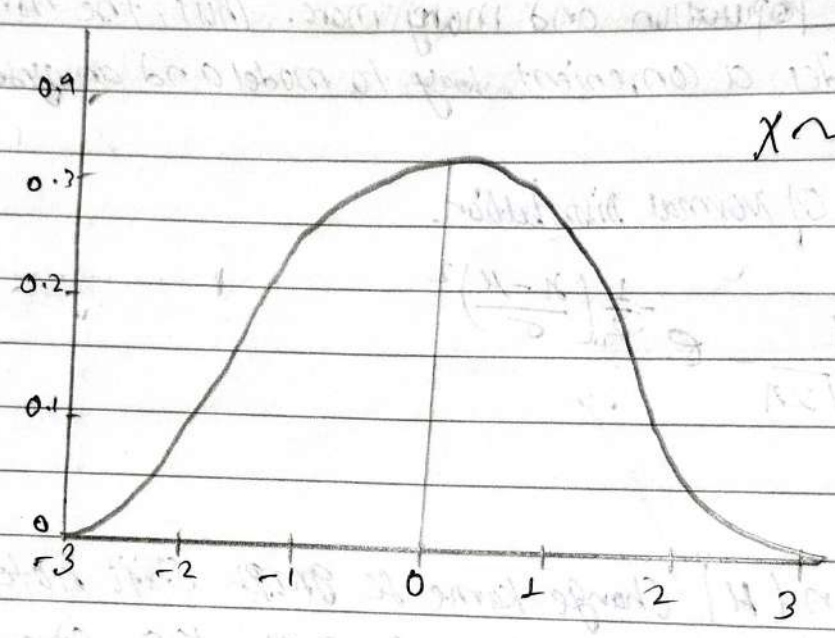
PDf Equation of Normal Distribution.

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

Note → • Mean (μ) Change karne se graph Shift hote hai
• Std (σ) Change karne se graph ka Spread Change hota hai

Standard Normal variate (Z)

↳ A standard Normal variate (Z) is a standardize form of the normal distribution with mean = 0 and std = 1.



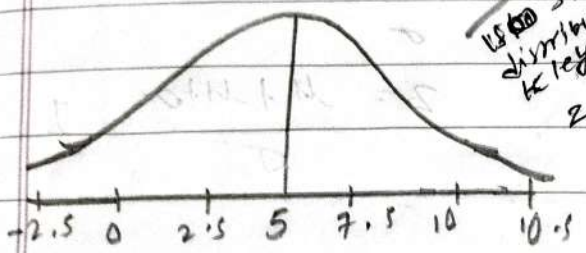
Standarizing a normal distribution allows you to compare different distribution with each other, and to calculate probability using standarized table or software.

Equation:

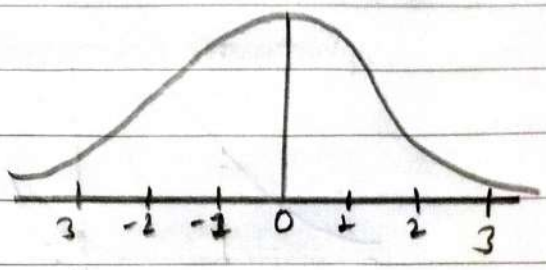
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- How to transform a normal distribution to standarize Normal variable.

$$Z = \frac{X - \mu}{\sigma}$$



Isko standard normal distribution banana ho to $Z = \frac{X-\mu}{\sigma}$ karo no jaega

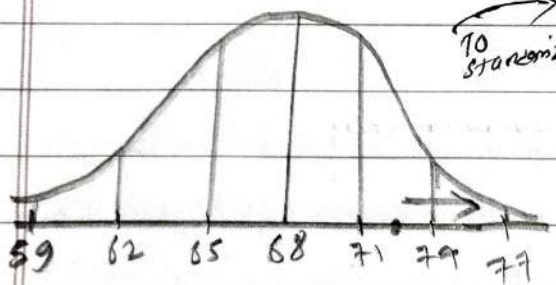


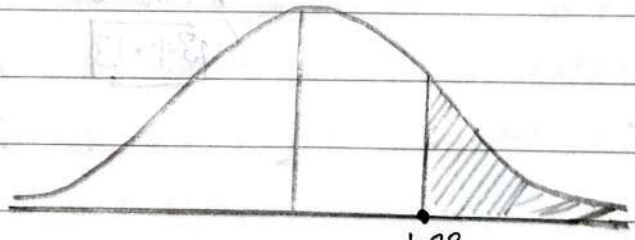
Kya faida hai karne ka?

Suppose the height of adult males in a certain population follow normal distribution with mean of 68 and a std of 3 inches. What is the Probability that a random seleted male from this population is taller than 72 inches.

$X \sim N(68, 3)$

To standardize → Z | $Z = \frac{72-68}{3} = \frac{4}{3} = 1.3$ ← Z score





Z-table

↳ A Z-table is used in Statistic to find the Probability that a standard normal variable (Z) falls below a given value.

↳ ISS standard normal variable ma 1.33 se kam aane ka probability 0.90824 hai

90.82 hone chaiye 1.33 se jada

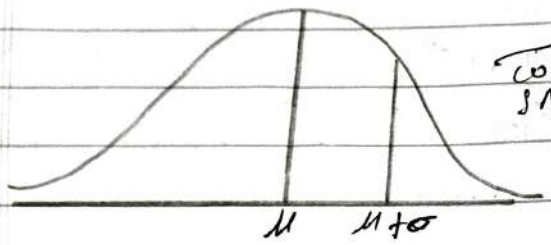
$100 - 90.82$

$= 9.1\%$

for a Normal Distribution $X \sim (\mu, \text{Std})$ What Percentage of Population lie between mean and 1 standard deviation, 2 std and 3 Std?

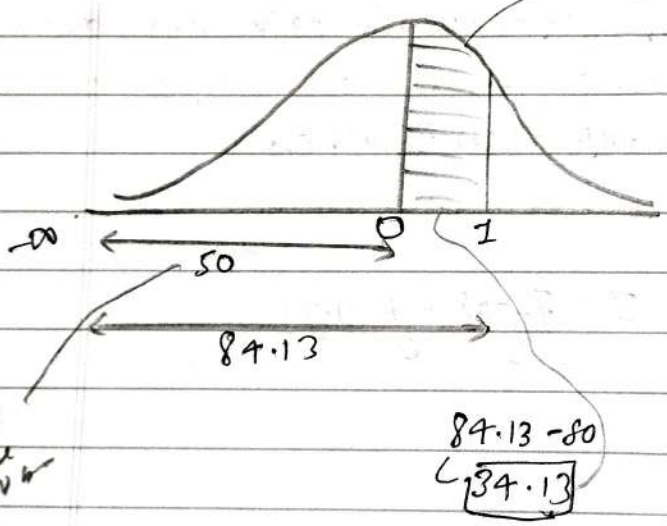
$X \sim N(\mu, \sigma)$

$$z = \frac{\mu - \mu}{\sigma} = 0$$

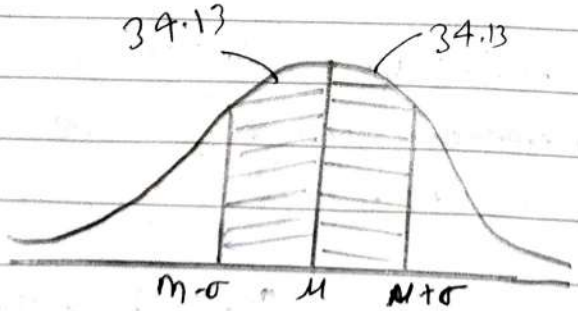


(convert SNV(Z))

$$z = \frac{\mu + \mu + \sigma}{\sigma} = 1$$



84.13 - 50
= 34.13



-1σ $+1\sigma$

64.2

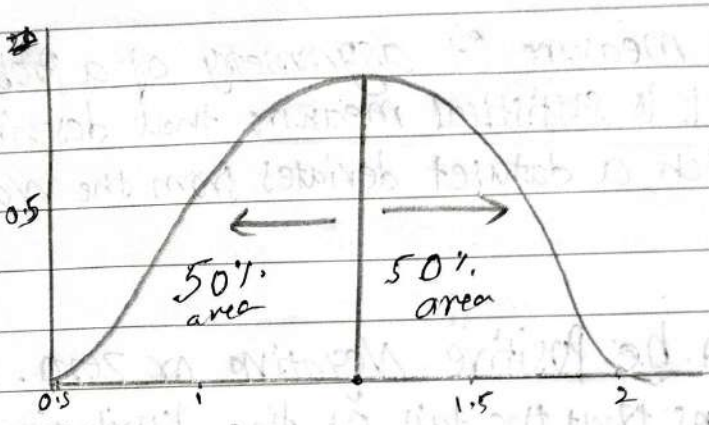
-1σ se $+1\sigma$ tak 64.2% Data adalah nilai

empirik

Properties of Normal Distribution

① Symmetricity

↳ The normal distribution is symmetric about it mean which means that the probability of observing a value above mean is the same as the probability of ~~observation~~ observing a value below the mean. The bell shaped curve of the normal distribution reflects this symmetry.

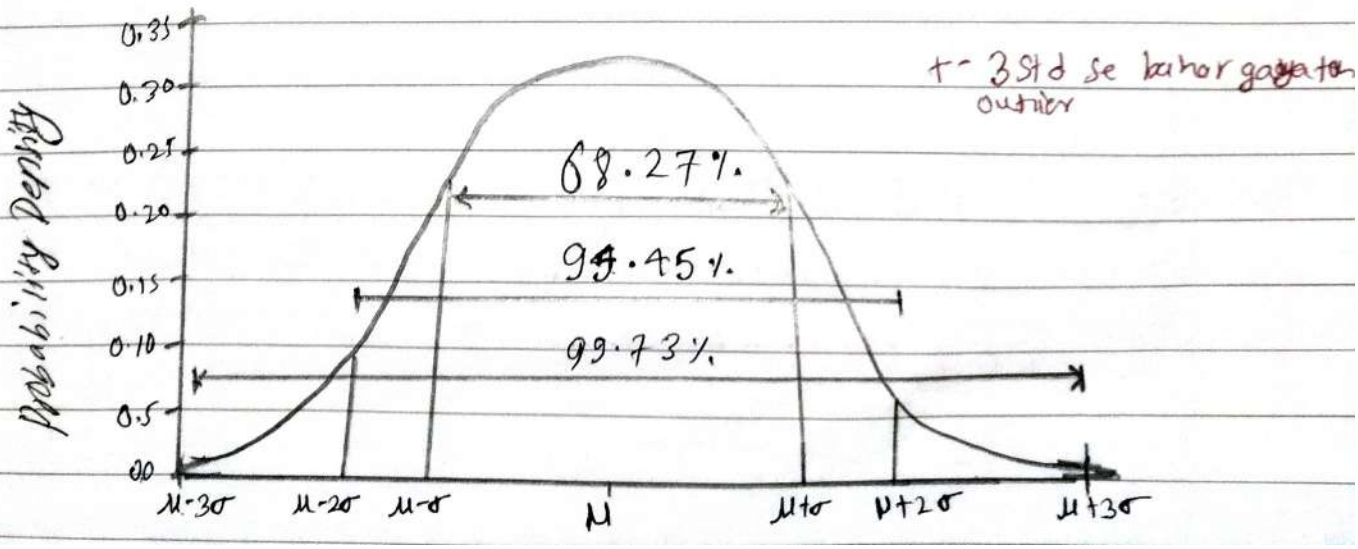


② Measure of central tendencies are equal → mean → mode → median.

③ The area under graph is 1.

④ Empirical rule

↳ The Normal distribution has a well know empirical rule, also called the 68-99-99.7 rule, which ~~states~~ that ~~Approximately~~ 68% of the data fall ~~within~~ one std of the mean, 95% of the data falls within 2 std of the mean, and about 99.7% of the data falls under the 3 std of the mean.



- What is Skewness ?

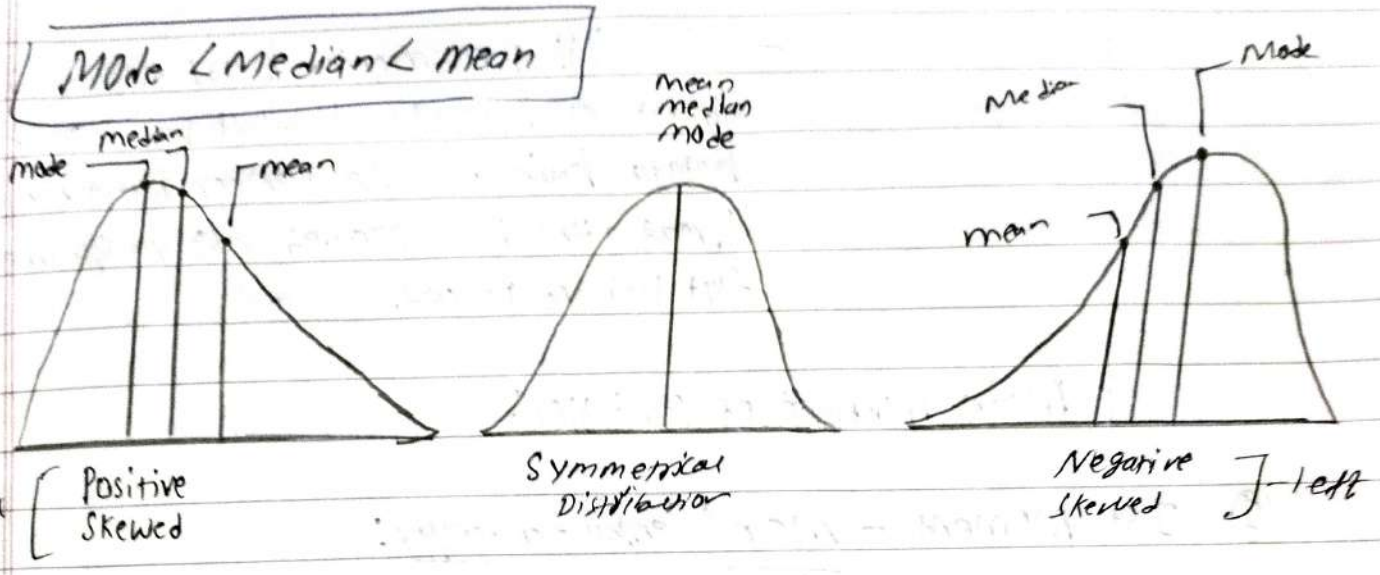
→ A Normal Distribution is a bell-shaped symmetrical distribution with a specific mathematical formula that describe how data is spread out. Skewness indicates that the data is not symmetrical, which means it is not normally Distributed.

→ Skewness is a measure of asymmetry of a probability distribution. It is statistical measure that describe the degree to which a dataset deviates from the Normal Distribution.

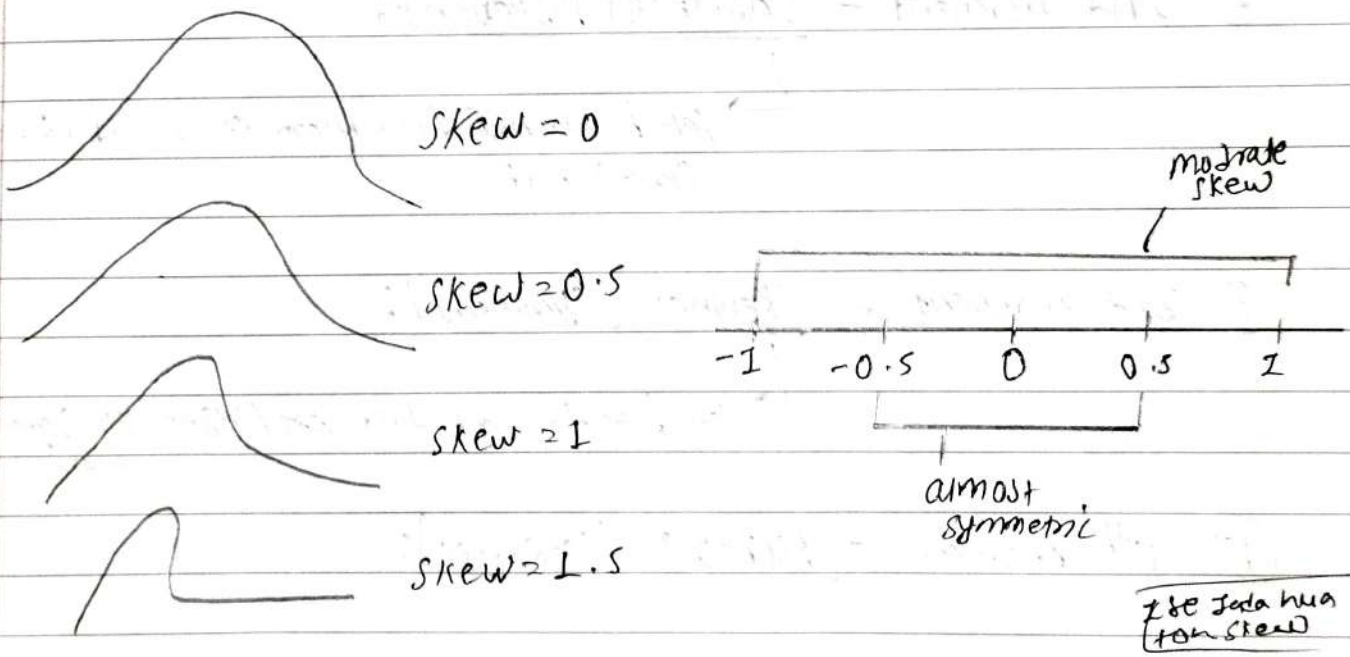
→ Skewness can be positive, Negative or zero. A positive skewness means that the tail of the distribution is longer on the right side, while Negative skewness means that the tail is longer on left side. A zero skewness indicates a perfect symmetrical distribution.

→ In a symmetric distribution mean, median, mode are equal. In contrast, in a skewed distribution, the mean, median, mode are not equal, and the distribution tends to have a longer tail on one side than the other.

Mode < Median < mean



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



Ise Jada hua toh skew

Use of Normal distribution

↳ Outliers detection

↳ Hypotesis testing

↳ Central limit theorem

↳ Assumption on data for ml algorithms

Note → agar mean - 3x std
or
mean + 3x std
ke ya toh outlier find kar sakte ho iss se

Stastical moments → Statistical moments basically wo quantities hain jo distribution ka shape batata hain - jaise center kahan hai, spread kitna hai, skewed hai ya symmetric, flat hai ya peaked.

4 Main moments of statistics:

① 1st moment - Mean (central tendency):

↳ Yeh batata hai data ka centre kahan hai

② 2nd moment - Variance (Dispersion):

↳ Yeh batata hai data mean ke around kitna spread hai

③ 3rd moment - Skewness (Symmetry):

↳ Yeh batata hai data left/right ya symmetric hai;

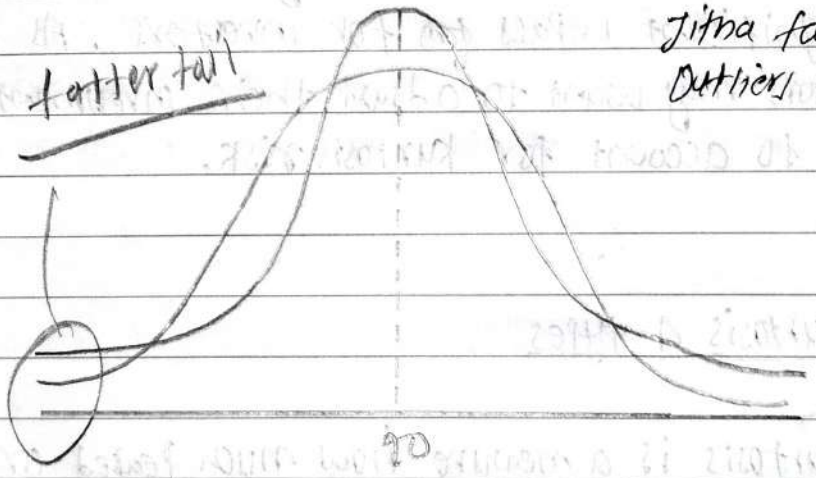
④ 4th moment - Kurtosis (Peakedness):

What is Kurtosis ?

↳ Kurtosis is the 4th stastical moment. In probability and stastics Kurtosis (meaning "curved, arching") is measure of the "tailedness" of the probability distribution of a real-valued random variable. Like skewness, kurtosis describe a particular aspect of a probability Distribution.

⇒ Data ke tails ma kitne heavy ya light (ie, extreme values kitne zada ya kam aate hain) yeh batate hai

fatness = Presence of outliers
jitna fatter tail utna zada Outliers hone ka Chances



formula simple kurtosis

$$\left\{ \frac{n*(n+1)}{(n-1)*(n-2)*(n-3)} * \sum_{I}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3*(n-1)^2}{(n-2)*(n-3)}$$

- Practical Use-Case

In finance, Kurtosis risk refers to the risk associated with the possibility of extreme outcome or "fat tails" in the distribution of returns of a particular asset or portfolio.

If a distribution has high kurtosis, it means that there is a higher likelihood of extreme events occurring, either positive or negative, compare to a normal distribution

In finance, kurtosis risk is important to consider because it indicates that there is greater probability of large losses or gains occuring, which can have significant impact for investors. As a result, the investors may want to adjust their investment strategies to account for kurtosis risk.

- Excess kurtosis 4 types

↳ Excess kurtosis is a measure how much peaked or flat distribution is compare to normal distribution, which is considered to have a kurtosis of 0. It is calculated by subtracting 3 from the sample kurtosis's coefficient.

In myword

Kurtosis batata hai ke tails kitna fat hai (kitne extreme value hai) distribution me
Lekin "excess kurtosis" sirf ek adjust kiya version hai kurtosis ka - matlab

$$\text{Excess kurtosis} = \text{kurtosis} - 3$$

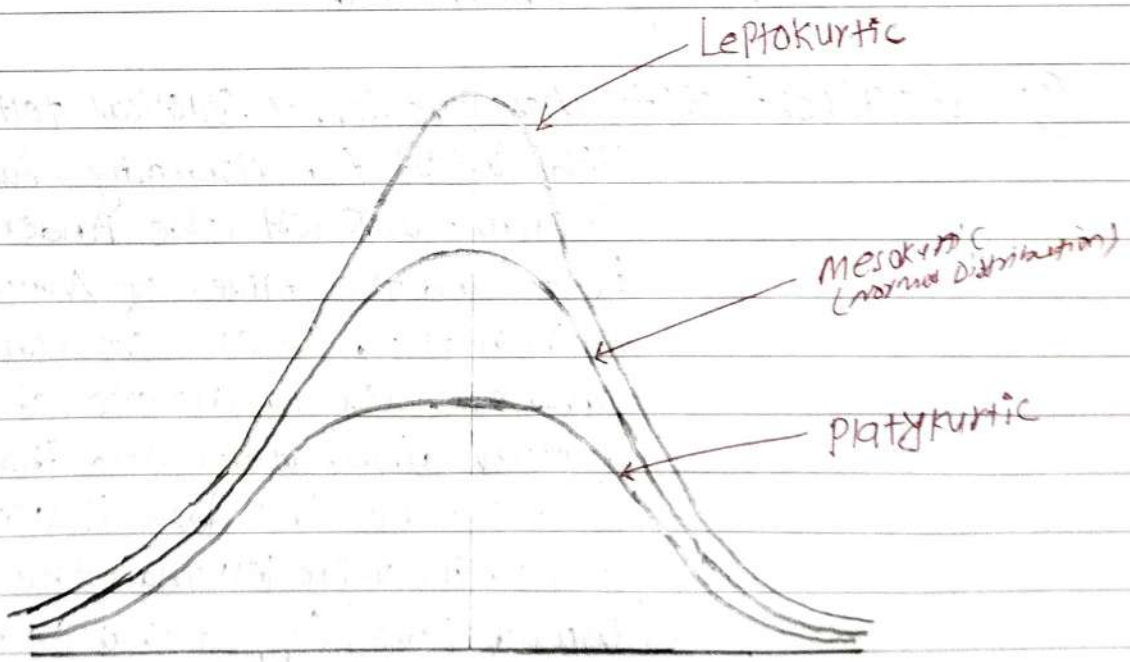
→ 3 minus karte hai q ke normal distribution ka kurtosis = 3 hota hai
→ Toh agar usme se 3 hata do, toh normal distribution ka excess kurtosis = 0 ban jata hai
- easy comparison k liye.

Types of Kurtosis

Types	Excess Kurtosis	Tail Shape	Meaning
Mesokurtic	0	Normal Jaisa	Normal distribution
Leptokurtic	>0	fat tails, sharp Peak	Outlier Zada
Platykurtic	<0	flat tails, bord shape	outliers Kam

→ Excess kurtosis tell how heavier or lighter the tails are compared to normal distribution

Agar +ve hai → tails heavy → Zyada extreme values.
Agar -ve hai → tails light → Kam extreme values.



- How to find if a given distribution is normal or not ?

① Visual Inspection → One of the easiest ways to check for Normality is to visually inspect a histogram or a density plot of the Data. A Normal distribution has bell shaped curve, which means that the majority of the data falls in the middle, and the tails tapper of symmetrically. If the distribution looks approximately bell-shaped, it is likely to be Normal

② QQ Plot → Another way to check for Normality is to create a Normal Probability Plot (also know as Q-Q plot of the data. A Normal probability plot Plots the observed data against the expected value of a Normal distribution. If the data points falls along a straight line; the distribution is likely to be normal.

③ Stastical tests → There are serval stastical test than can be used to test for normality, such as the Shaipro-Wilk test, the Anderson-Darling test, and the ~~values of Normal~~ Kolmogorov-Smirnov test. These test Compare that Observed data to the expected values of a Normal distribution and provide a p-value that indicates the data is likely to be normal or not. A P-value less than the significance level (usually 0.05) suggest that the normal data is Normal.

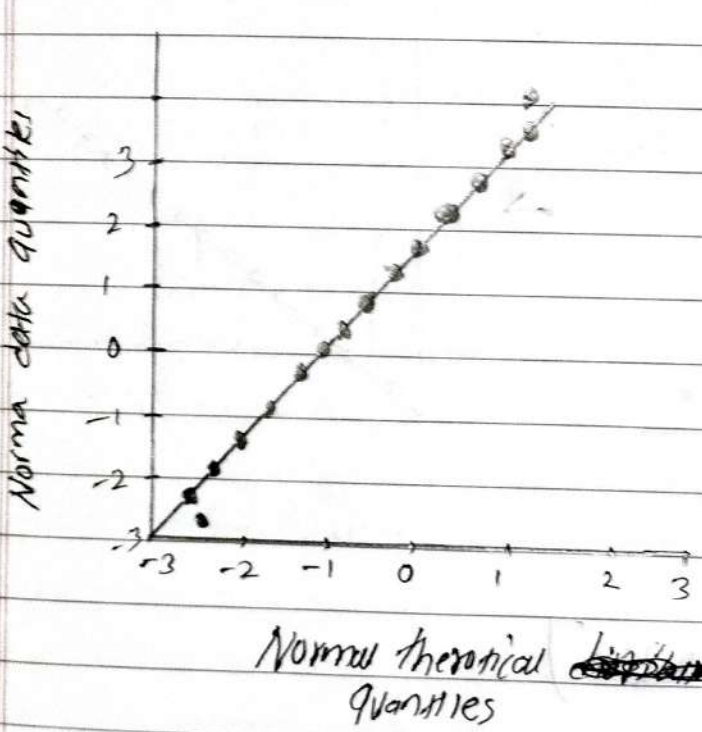
- What is a QQ Plot and how it is plotted?

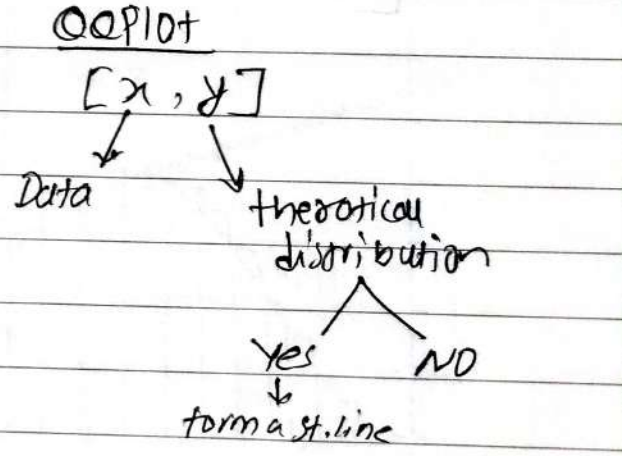
↳ A Q-Q Plot (Quantile - Quantile plot) is a graphical tool to check wheather your data follows a specific theoretically distribution - usually the normal.

Concept

→ You plot your data quantiles (sorted values) against the quantiles of theoritical distribution (eg.. normal).

→ If your data actually follows that distribution, the points will fall roughly on straight line (x=y).



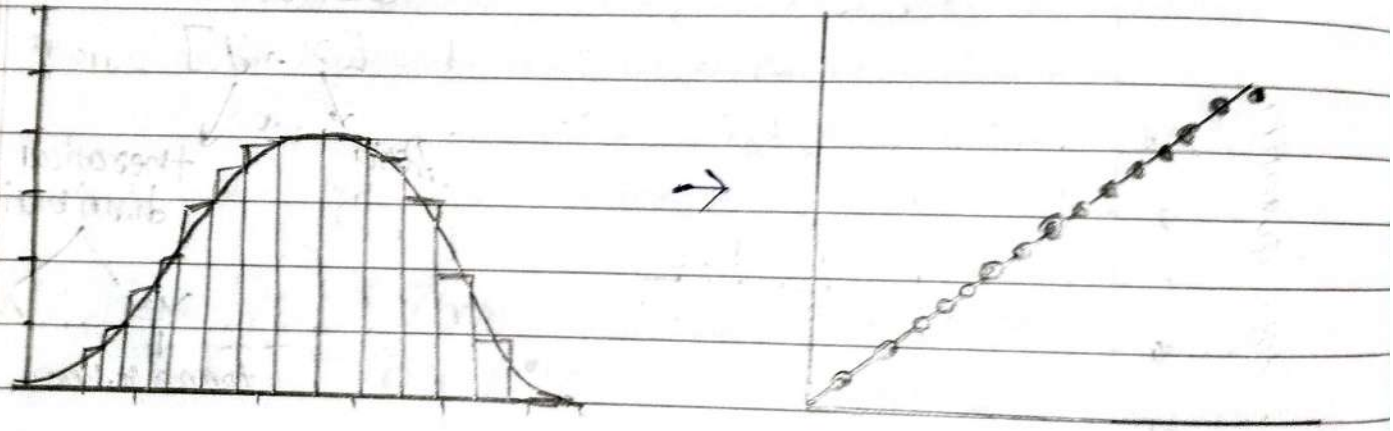


→ QQ Plot batata hai ke tumhare data kis distribution k jaise dikhta hai - agar seedhi line hai, sab set hai;
Agar tedha hai, kuch gadbad hai.

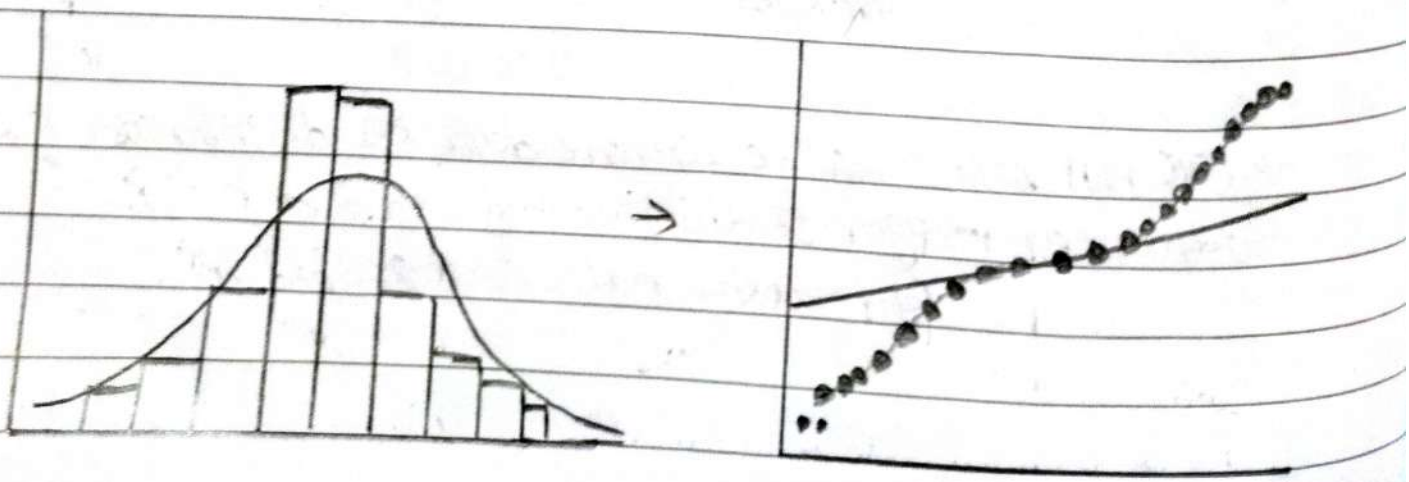
→ In QQPlot, the quantiles of two set of data are plotted against each other. The quantiles of one set of data are plotted on X-axis, while the quantiles of the other set of data plotted on the y-axis. If two set have the same distribution, the QQPlot will fall on a straight or straight line. If two sets of data do not have the same distribution, the points will deviate from the straight line.

- HOW to interpreate QQ Plot.

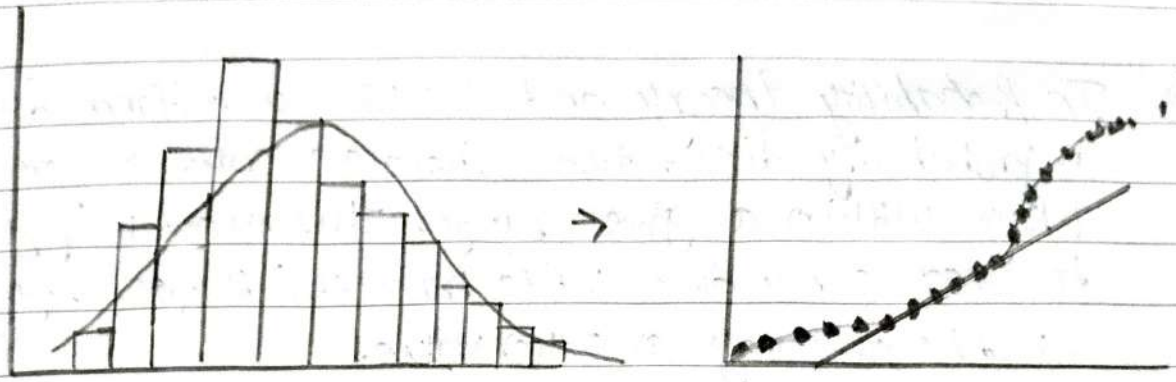
[Normal Distribution Data]



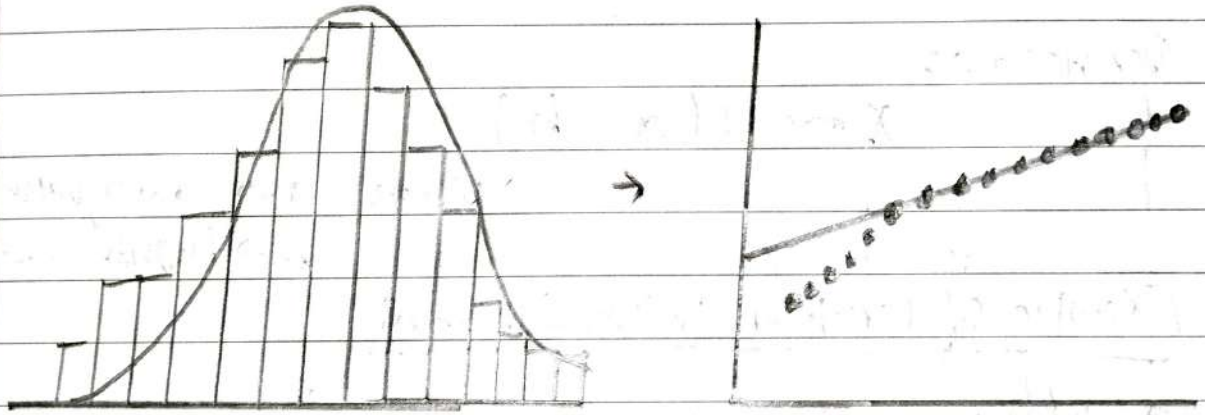
[Data too Peaked in middle.]



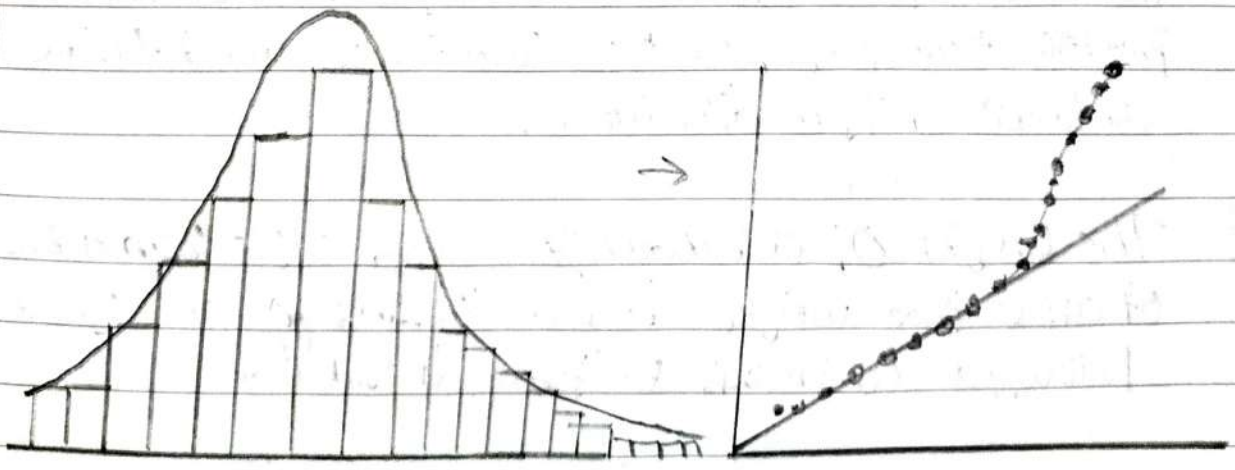
[Skewed Data]



[Skewed left]



[Skewed right]



What is Uniform Distribution

↳ In Probability theory and stastics, a uniform distribution is Probability distribution where all outcomes are equally likely within a given range. This means if you were to select a random value from this range, any value would be as likely as other value.

Types — Uniform distribution

- Discrete
- Continous

Denotes as:

$$X \sim U(a, b)$$

where $a \rightarrow$ lower value
$b \rightarrow$ higher value

Example of Continous Uniform distribution:

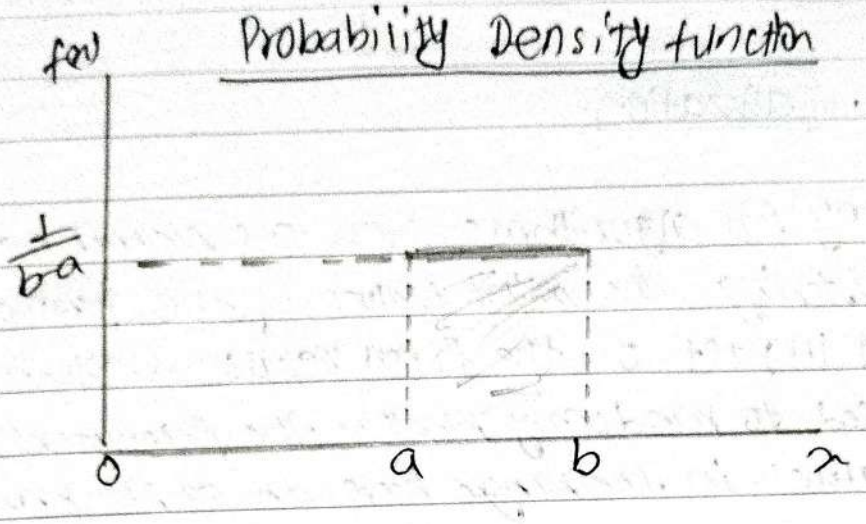
(1) The height of a person randomly selected from a group of individuals whose heights range from 5'6" to 6'0" would follow a Continous Uniform distribution.

(2) The time it takes for a machine to produce a product, where the production time range from 5 to 10 minutes, would follow a Continous Uniform distribution.

(3) The weight of a randomly selected apple from a basket of apple that weights between 100 and 200 gm, would follow a continous Uniform distribution.

PDF CDF and graphs.

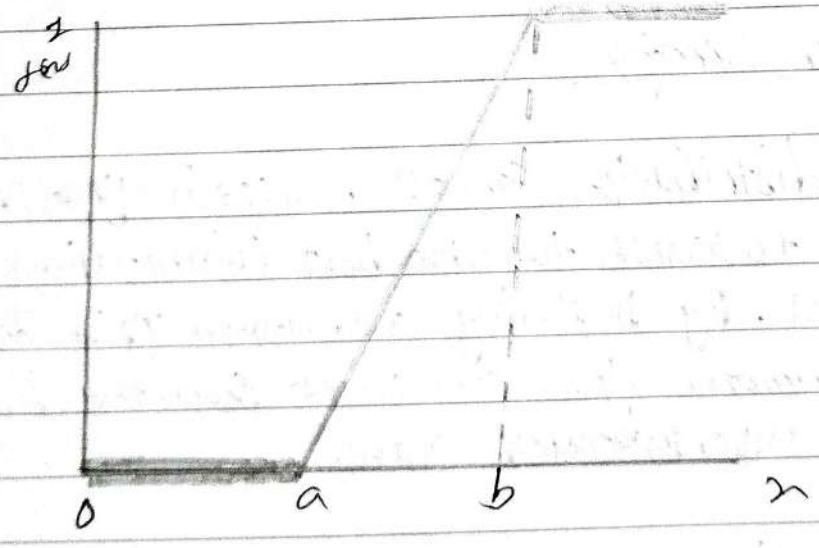
Probability Density function



Uniform equation

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le X \le b \\ 0 & \text{for otherwise} \end{cases}$$

Cumulative distribution function



Skewness → 0 → symmetric → normal.

Uniform distribution

- Application in machine learning and Data science.

(a) Random initalization

↳ In many ML algorithms, such as Neural networks and K-mean clustering, the inital values of the parameters can have a significant impact on the final result. Uniform distribution is often used to randomly initalize the parameters, as it ensures that all values in the range have an equal probability of being selected.

(b) Sampling

↳ Uniform distribution can be used for sampling. for If you have a dataset with an equal no of samples from each class, you can use uniform distribution to randomly select a subset of the data that is representative of all the classes.

(c) Hyperparameter turning

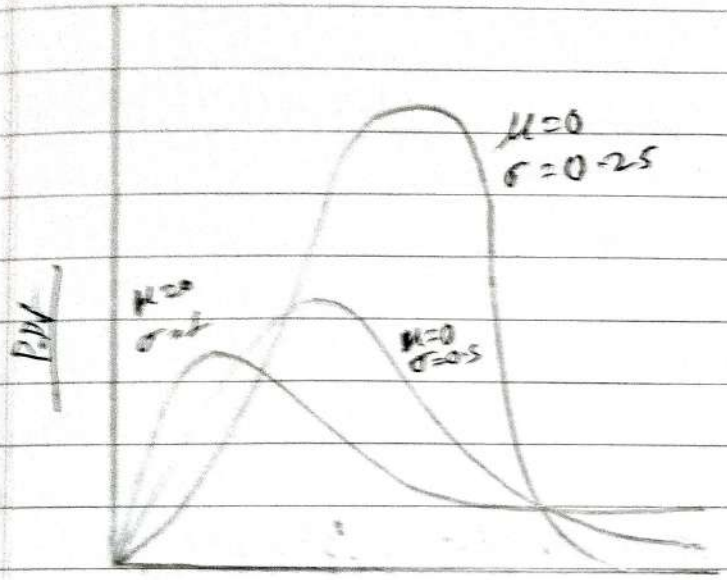
↳ Uniform distribution can also be used in Hyperparameter turning where you need to search for the best combination of Hyperparameters for a ML model. By defining a uniform prior distribution for each Hyperparameter, you can sample from the distribution to explore the Hyperparameter space.

(d) Data augmentation

↳ In some case, you may want to artificially increase the size of your dataset by adding new examples that are similar to original data. Uniform distribution can be used to generate new Data Points that are within a specificied range of the original data.

Log Normal Distribution

> In probability theory and statistics, a lognormal distribution is heavily tailed continous probability Distribution of a random variable whose logarithm is normally Distributed



Examples

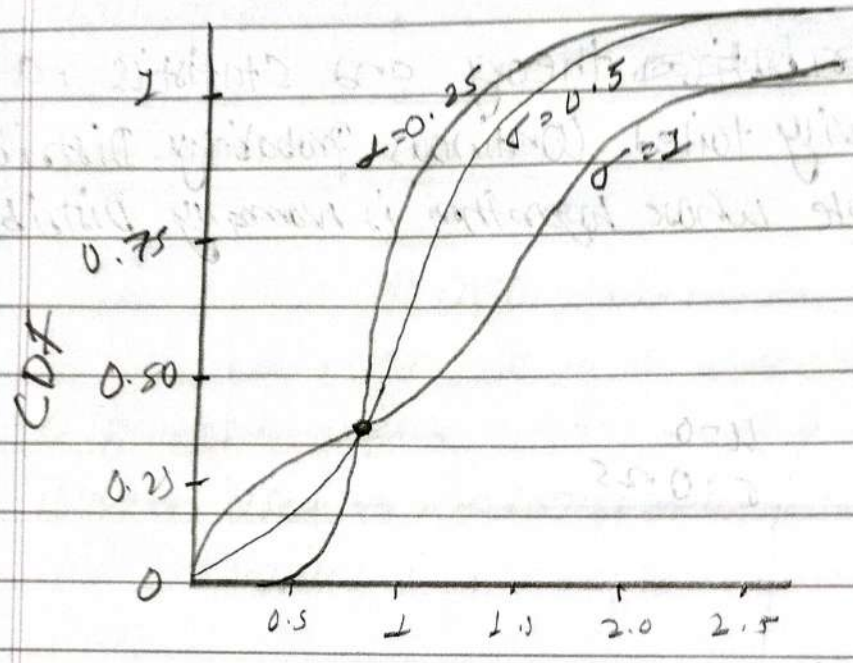
- The lenght of Comment Posted in internet discussion forms follow a log-normal distribution
- Users dwell time on online article follow a log normal distribution
- The lenght of Chess game tends to follows a lognormal distribution
- In economic, There is evidence that the income of 97%-98% of the Population is distributed log normaly.

Denoted by: $X \sim \text{lognormal}(\mu, \sigma)$

PDF equation

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}, \quad \text{for } 0 < x < \infty$$

CDF



Skewness → Skewed

- How to check a random variable is log normally Distributed?

If x is log normally Distributed, then $y = \log(x)$ should be normally distributed. to check Normal distribution of y Plot a qqplot if straight line banta hai then confirm log normally Distributed hai

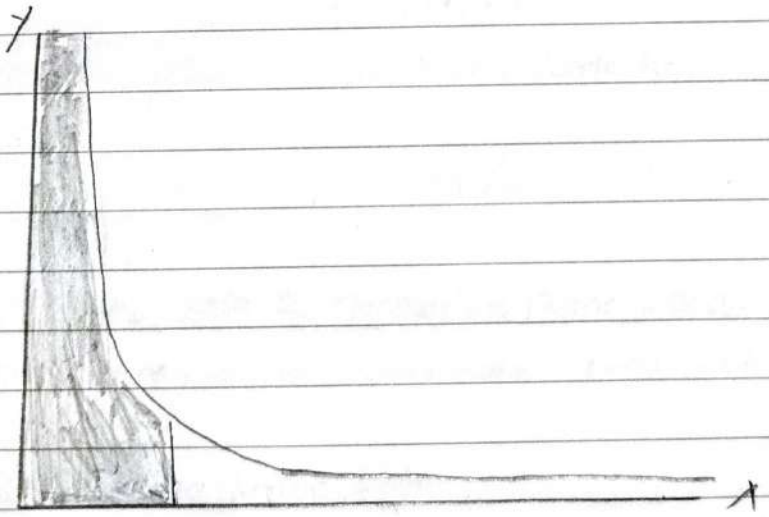
Pareto Distribution

↳ The Pareto Distribution is a type of Probability Distribution that is commonly used to model the distribution of Wealth, Income and other quantities that exhibit a similar Power-law behaviour

What is Power law

→ In mathematics, a power law is a functional relationship between two variable, where one variable is proportional to Power of the other. Specifically, If y and x are two variables related by a Power law, then the relationship can be written as:

$$Y = K * X^a$$



Vilfredo Pareto originally used this distribution to describe the allocation of wealth among individuals since it seemed to show well the way that larger portion of the wealth of society is owned by smaller % of the people in the society. He also Used it to describe distribution of income. The idea is something expressed more simply as the Pareto principle or the "80-20" rule which says that 20% of the Population control 80% of the wealth.

Mathematical transformation

- Mathematical transformation like log, sqrt, reciprocal, Power, Box-cox, Yeo-Johnson can make a distribution more normal, if it is skewed.

But yeh guarantee nhi hai ke har transformation se distribution Perfect Normal ho jaye. Kuch distribution extreme hai (bimodal, heavy tail), unko normal banna impossible hai with just a simple transform.

Log transform

↳ log transform ek mathematical transformation hai jisme ham kisi variable x ka logarithm leta hai:

$$Y = \log(x)$$

- Sirf Positive values k liye kaam karta hai

Use kaha hota hai / Kyu karte hai

left k leye square use karo

① Right-skewed data ko normalize karne k leye:
↳ ex income, population, sales data - bahut bade values ko compress kar deta hai

② Variance Stabilize karne k leye
↳ Jab high values spread bohot jada hota hai, log transform se variance relatively stable ho jate hai

③ Outliers ko reduce karne k leye
↳ Extreme value ko compress karke analysis aur modling ma influence kam karte hai.

Discrete Distribution

Bernoulli Distribution

↳ Bernoulli Distribution ek discrete Probability distribution hai jisme sirf do possible outcome hote hei:
Success (1)
failure (0)

Example: Cointoss → Head (1), tail (0)
Product test → Pass (1), fail (0)
Spam detection → Spam (1), Not spam (0)

P → Probability of success (1)
P-1 → Probability of failure (0)

Only 1 Parameter P

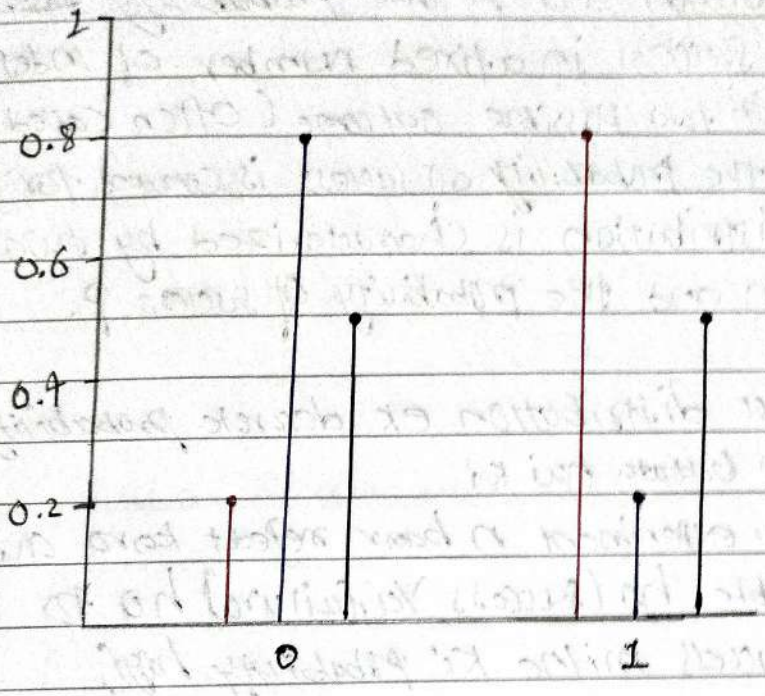
PMf :

$$P(X=x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

P(X=x) batata hai ke success [1] ya failure [0] hone ki kitne chance hai

The Bernoulli distribution is commonly used in ml for modeling binary outcomes, such as Whether a customer will make a Purchase or not, whether a Patient will have a certain disease or NOT.

PMF of Bernoulli distribution



Three example 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 $P(x=1) = 0.5$

Binomial Distribution

↳ Binomial Distribution is a Descrete Probability Distribution that describe number of success in a fixed number of independent Bernoulli trails with two possible outcome (often called "success" and failure), where the probability of success is constant for each trail. The binomial distribution is characterized by two Parameters the no of trials n and the Probability of success P.

Note → Binomial distribution ek descrete probability distribution hai jo batata hai ki agar koi experiment n baar repeat karo aur 1 baar me do possible ho (success ya failure) ho to kitni baar success milne ki probability hogi.

The Probability of anyone watching this lecture in future and liking is 0.5. What is the probability that:

(1) NO one out of 3 People will like it.
↳ $\frac{1}{8}$

(2) 1 out of 3 People will like it.
↳ $\frac{3}{8}$

(3) 2 out of 3 People will like it.
↳ 3/8

YYY
YYN
YNY
YNN
NYY
NYN
NNY
NNN
} 8

have 2

Binomial distribution ^Parameter :

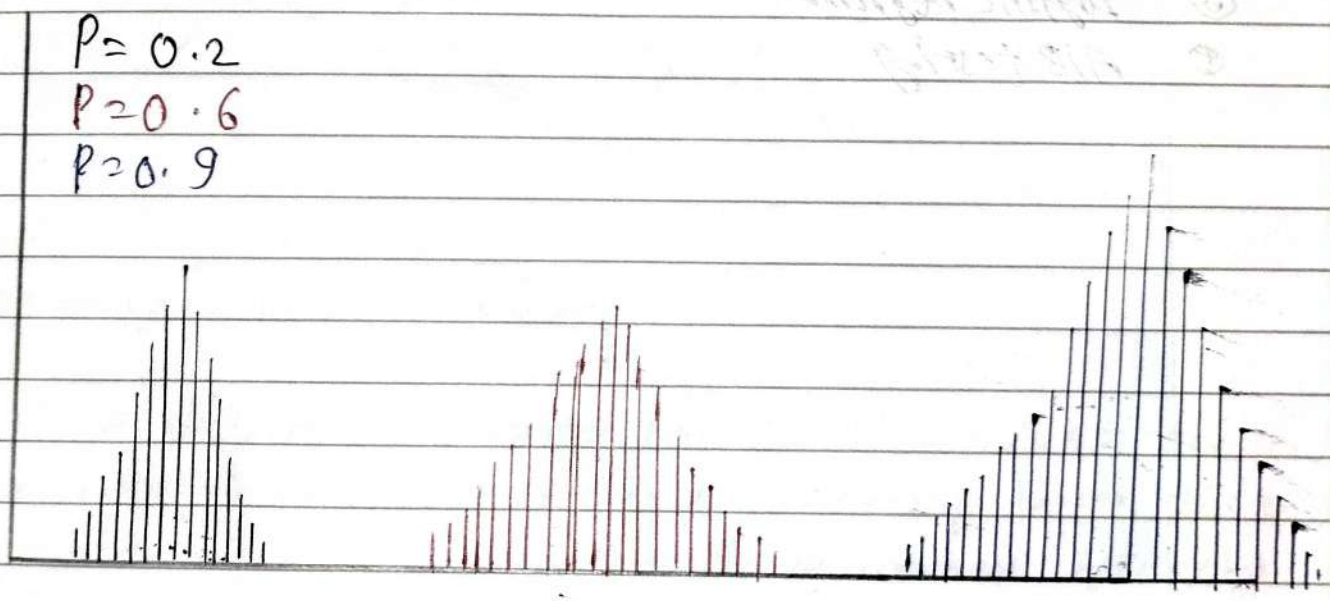
n → no of trials

P → probability of success.

PMF of Binomial distribution :

$$P(X=K) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial distribution with Different Probability of success



NOTE

Jab P = 0.5 distribution Perfect symmetric hoti hai - bell ma centred.

Jab P < 0.5, distribution right skewed (tail right side pe hoti hai)

Jab P > 0.5, distribution left skewed (tail left side pe hoti hai)

Certain:

1. The process consist of n trials.
2. only 2 exclusive outcome are possible, a success and a failure
3. P(success) = p and P(failure) = 1-p and it is fixed from trail to trial
4. The trails are independent.

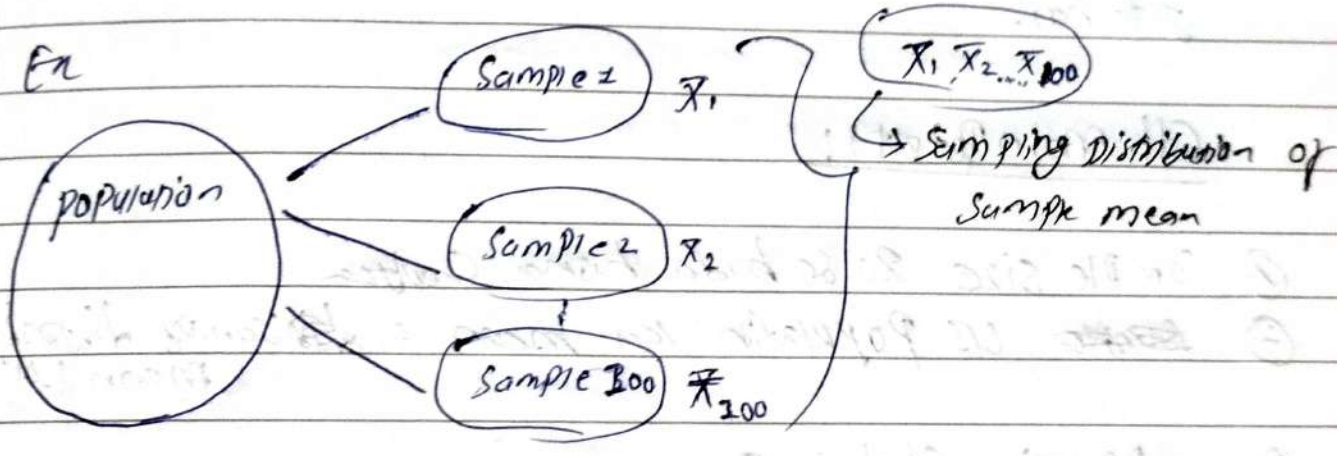
Where Binomial distribution used in:

1. Binary classification problems
2. Hypothesis testing
3. Logistic regression
4. A/B testing.

Sampling Distribution

↳ Sampling distribution ka matlab hai, jab population se equal size ~~of sample~~ k multiple sample lena k baad, un sab samples ka mean (ya koi aur statistic) ka set jo banta hai, wahi hota hai sampling distribution.

Ex



Why Sampling Distribution is important?

↳ Sampling distribution important hai kyunki yeh batata hai ke sample statistic population k respect ma kitne vary karega, jiss se hum confidence intervals, hypothesis tests, aur prediction kar sakte hai.

mean median mode

Central Limit therom

→ Agar hum kisi bhi population (chahe wo normal ho ya na ho) se bohat saare random samples lein. aur hum har sample ka mean nikale, toh un sub sample means ka distribution (graph) normal ban jata hai.

Clt main Points:

① Sample size 30 se bada hona chahiye

② Us Population ka mean = sample distribution mean (μ)

③ Population std = σ
Sample size = n

$$\text{Sample distribution} = \frac{\sigma}{\sqrt{n}}$$

Sample mean = POPulation mean = μ

$$\text{Sample Std} = \frac{\sigma}{\sqrt{n}}$$

→ The Clt is important in statistic and ml because it allows us to make probabilistic inferences about population based on sample of data. for ex: we can use Clt to construct confidence intervals, perform hypothesis tests, and make predictions about the mean based on the sample.

Some terms

Population → A Population is the entire group or set of individuals, objects, or events that a researcher want to study or draw conclusion about it.

Sample → A sample is a subset of the population that is selected for study.

Parameter → Parameter ek number hota hai jo population ki property batata hai like, mean, std, var etc.

Statistic → ~~Sample ek~~ statistic ek number hai jo sample ki property batata hai like μ, σ, ~~var~~ etc.

Population → Parameter → Mostly unknow
Sample → statistic → Calculate kar sakte ho

Point estimate → ek single Numerical value jo population Parameter ka best guess batata hai, sample ke basis pe.

In other words: mean var

Sample ke statistic ek Numerical value hota hai: Jab ham us value ko use population ka parameter ka aandaza lagane k liye ~~kabo~~ keye jata hai, tab use Point estimate kahete hai.

Confidence interval (CI)

Confidence interval ek range of values hai jisme Population Parameter (like mean, Proportion etc) hone ki Probability hoti hote hai - based on sample data

- Sample ke data se Confidence interval nikalte hai, aur uss range se Population ka asli value guess karte hai.

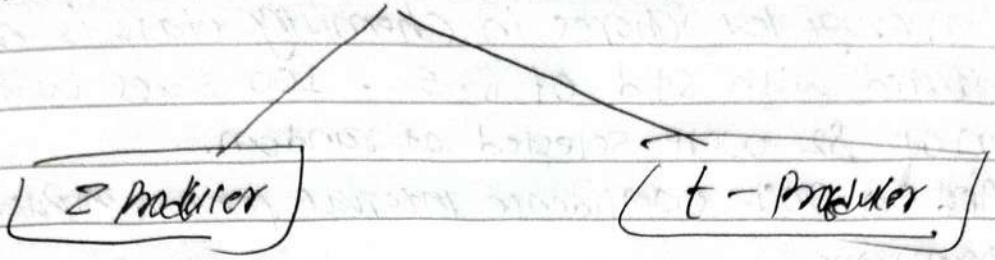
→ Confidence interval is created for Parameters and NOT for statistics. statistics helps us get the Confidence interval for a parameter.

Confidence level

CI usually expressed as a Percentage like 95%, Indicates how sure we are that the true values lies within the interval.

Confidence interval = Point estimate ± Margin of error

Ways to calculate confidence interval:



① Z Producer (sigma known)

Assumption

ⓐ Random sampling → The data must be randomly selected to avoid bias

ⓑ Known population standard deviation

ⓒ Normal Distribution or large sample ($n \geq 30$) → The Population should be normal, or the sample should be large enough for the central limit theorem to apply.

formula of CI using Z producer:

$$CI = \bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

$\bar{X}$ = sample mean
σ = population standard deviation (known)
n = sample size
$Z_{\alpha/2}$ → critical value (from z-table)
90% CL → 1.645
95% CL → 1.96
99% CL → 2.575

Z can calculate by Area to left $= \frac{(CI + I)}{2}$

Put this value in place of $Z_{\alpha/2}$

(1) The average test scores in Chemistry class is normally Distributed with std of 6.5 . 100 scores with sample mean of 82 were selected at random.

(a) find a 90% confidence interval for the population mean exam score.

(b) find the value of margin of error

soln,

$\sigma = 6.5$, $n = 100$, $\bar{x} = 82$

Area to left $= \frac{CL + 1}{2} = \frac{0.90 + 1}{2} = \boxed{0.95}$

$Z_{\frac{\alpha}{2}} \rightarrow 1.645$

$CI = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$= 82 \pm 1.645 \left(\frac{6.5}{\sqrt{100}} \right)$

$= 82 \pm 1.06925$

$\boxed{[80.93075 , 83.06925)}$

Margin of error $= Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$

$= 1.645 \left(\frac{6.5}{\sqrt{100}} \right)$

$= \boxed{1.06925}$

Interpreting confidence interval?

A confidence interval is a range of values within a population parameter such as the population mean, is estimated to lie within a certain level of confidence. The confidence interval provides an indication of the precision and uncertainty associated with the estimate. To interpret the confidence interval values, consider the following points:

① Confidence level

↳ The confidence level (commonly set at 90%, 95% or 99%.) represents the probability that the confidence interval will contain the true population parameter if the sampling and estimation process were repeated multiple times. For ex, a 95% confidence interval means that if you were to draw 100 different samples from the population and calculate the confidence interval for each, approximately 95 of those interval would contain the true population parameter.

② Interval range

↳ The width of the confidence interval gives an indication of the precision of the estimate. A narrower confidence interval suggest a more precise estimate of the population parameter, while wider interval indicate greater uncertainty. The width of the interval depends on the sample size, variability in the data, as the desired level of confidence.

95% ↑ → width ↑

narrow - chotta
wide - bada

factor that effect confidence interval.

Sample size

↳ larger sample size - narrow confidence interval

↳ more data give more precise estimate of the population parameter

ex

$n = 30 \rightarrow CI(48.5, 51.5)$

$n = 300 \rightarrow CI(49.5, 50.5)$

Confidence level

↳ Higher confidence level → wider confidence interval

↳ To be more confident, you must allow a bigger range.

ex

90 % CL → narrow

99 % CL → wider

Note

100% CL = $-\infty$ to $+\infty$

Population variability (σ)

↳ Higher variability → wider CI

↳ more variation in data = less precision in estimation the mean

ex

$\sigma = 5 \rightarrow$ narrow CI

$\sigma = 20 \rightarrow$ wide CI

Confidence interval (sigma not known)

② T Procedure (sigma not known)

Assumption

① Random Sampling → The Data must be randomly selected to avoid Bias.

② Sample Standard deviation → The population std is unknow, and the sample std is used as an estimate.

③ Approximately normal Distribution → The t procedure assume that the underlying distribution is approx normally distributed, or sample size large enough to apply (CLT).

④ Independent observations → The observation in the sample should be independent of each other.

formula of CI using t procedure.

$$CI = \bar{X} \pm t_{\alpha/2, df} \frac{S}{\sqrt{n}}$$

sample std ka mean

kisne CL k leya chahiya aur check kro t table ma uska plate kya hai aur value nikalo.

$\bar{X}$ = sample mean
S = sample std ka mean
$\sqrt{n}$ → size of sample
df (No of degree) → n − 1
$t_{\alpha/2}$ → CL (Confidence Level)

Note → agar CI nikalne hai using t procedure to pehle sample mean nikalo fir sample std ka mean nikalo usko sample std k place ma lagao use bad t-table ma kisna confidence level k leya chahiye usne k leya value dekho aur df ma sample se kam ka value dekho.

- Hypothesis Testing kyu karte hai:

↳ Ham data se decision lena chahte hain - lekin har baar jo Difference ya Pattern data mein dikhta hai, wo real bhi ho sakte hai ya chance (randomness) se bhi ho sakta hai. Toh Hypothesis Testing se hum check karte hain ke:

Kya jo result mila hai actual meaningful hai ya sirf luck hai?

Hypothesis Testing

↳ Hypothesis Testing ek statistical method hai jis se hum yeh decide karte hai ke sample data ke basic par population ke bare main jo hum assume kar rahe hai wo sahi hai ya galat.

Real life example:

Maan lo tu ek medicine test kar raha hai:

Ab tu dekhna chahta hai ke nayi dawa sach main behetar kaam kar rahi hai ya nahi.
Tu data collect karega aur Hypothesis Testing se decide karega ki kya difference statistically significant hai ya sirf randomness (chance) se hua hai.

① Null Hypothesis (H0):

→ In simple terms, the Null Hypothesis is a statement that assume there is no significant effect or relationship between the variable being studied. It serves as the starting point for hypothesis testing and represents the status quo or the assumption of no effect until proven otherwise. The purpose of hypothesis testing is to gather evidence (data) to either reject or fail the null hypothesis in favour of the alternative hypothesis, which claims there is a significant effect or relationship.

Null Hypothesis → kuch bhi change/relationship/effect nhi hai
matlab default assumption : sab kuch normal, koi diff nahi

② Alternate hypothesis (H1 or Ha):

↳ The alternate Hypothesis, is a statement that contradicts the null hypothesis and claims there is a significant effect or relationship between the variable being studied. It represents the research hypothesis or the claim the researcher wants to support through statistical analysis.

Alternate Hypothesis → Alternate hypothesis bolta hai difference, relationship, or effect hai population me.

Important Points

- How to decide what will be Null hypothesis and what will be Alternate hypothesis (Typically the Null hypothesis says Nothing new is happening).
- We try to gather evidence to reject the Null hypothesis.
- It is important to note that failing to reject the Null hypothesis doesn't necessity mean that the Null hypothesis is true; it just means that there isn't enough evidence to support the alternative hypothesis.
- Hypothesis test are similar to jury trials, in a sense. In a jury trial, Ho is similar to the Not guilty verdict, and Ha is the guilty verdict. You assume in a jury that Defendant isn't guilty unless the Prosecution can show beyond a reasonable doubt that he or she is guilty. If the jury says the evidence is beyond a resonable doubt, they reject Ho, Not guilty, in favour of Ha, guilty.

Steps involved in Hypothesis Testing.

Rejection Region Approach

1. Formulate a Null and Alternate Approach
2. Select significant level → This is the probability of rejecting the Null Hypothesis when it is actually true, usually set (0.05 or 0.01)

3. Check assumption (example distribution)
4. Decide which test is appropriate (z-test, T-test, Chi-square, Anova)
5. State the relevent test statistic
6. Conduct the test
7. Reject or not reject the Null Hypothesis
8. Interpreate the result.

Performing a Z-test example 1

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average production was 50 units per day with a known population std of 5 units. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average production of 53 units per day. The company wants to know if the new training program has significantly increased productivity.

soln,

given, $\mu = 50$, $\sigma = 5$
$n = 30$, $\bar{x} = 53$

① $H_0 : \mu = 50$, $H_a : \mu > 50$

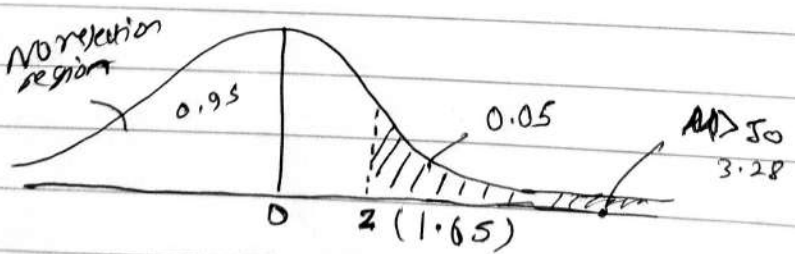
② $\alpha = 0.05 \rightarrow 5\%$

③ normality valid / pop std (σ) known

④ Z-test

⑤ Z

⑥ $$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{53-50}{5/\sqrt{30}} = \frac{3}{5/\sqrt{30}} = 3.28$$



⑦ We can reject the null hypothesis.

$\mu > 50$ will be

Example 2

Suppose a snack food company claims that their lays water packet contain an average weight of 50gm per packet. To verify this claim, a consumer ~~watching~~ watchdog organization decide to test a random sample of lays water packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50gms. The organization collects a random sample of 40 lays water packet and measure their weights. They find that the sample has an average weight of 49gms. with a known population std of 4gms.

Soln,

$$\mu = 50,\ n = 40,\ \bar{x} = 49,\ \sigma = 4$$

① $H_0 : \mu = 50$, $H_a : \mu \neq 50$

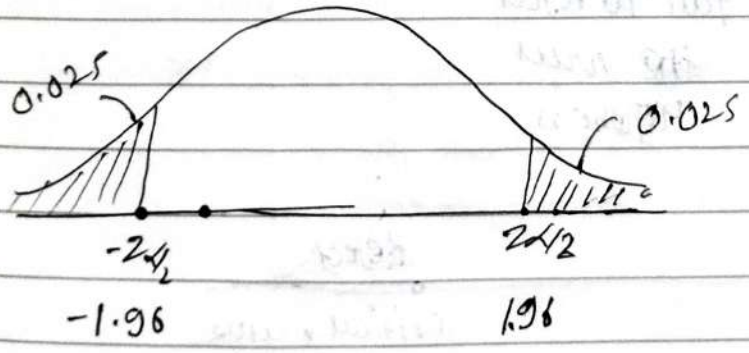
② $\alpha = 0.05$

③ Normality valid / σ known

④ Z test

⑤ Z

⑥ $$Z = \frac{49 - 50}{4/\sqrt{40}} = \frac{-1}{4/\sqrt{40}} = -1.58$$

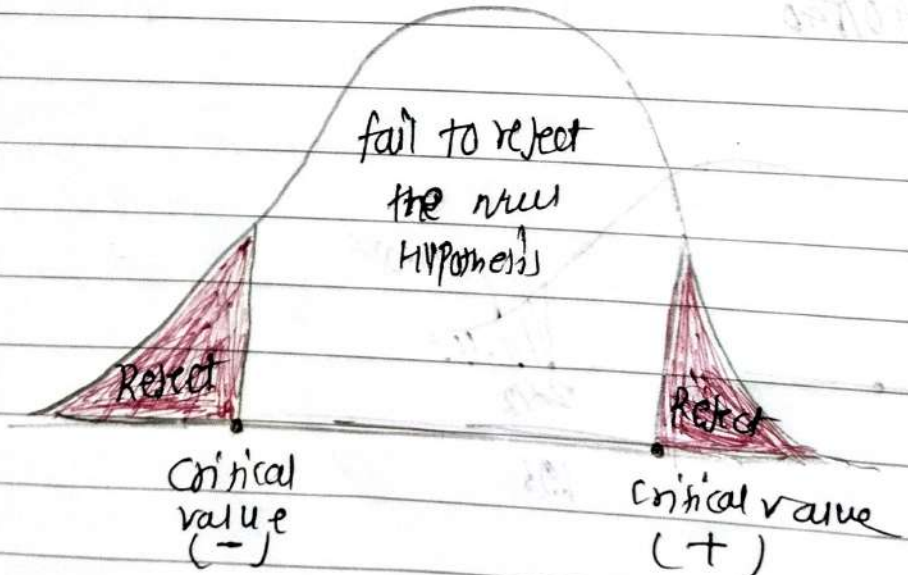


We can reject Null Hypothesis

Significance level → denoted by alpha (α), is a predetermined threshold used in hypothesis testing to determine whether the null hypothesis should be rejected or not. It represents the probability of rejecting the null hypothesis when it is actually true, also know as type I error.

Rejection Region → The rejection region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.

→ Rejection region (ya critical region) wo area hota jaha null hypothesis (H_0) ko reject kr dete hain.



Problem with rejection region approach

① result depend on the choosen significant level.
② Yeh sirf 'reject' ya 'accept' ka decision deta hai, evidence ka strength nhi deta hai
③ sample size se heavily effect hote hai

Type I vs Type 2 error

↳ In Hypothesis testing, there are two types of error that can occur when making a decision about the null hypothesis.

Type I (false Negative

↳ Type I error occur when the sample result lead to rejection of the null hypothesis when it is infact true.

In other words, it's the Mistake of finding a significant effect or relationship when there is none. The probability of committing a type I error is denoted by alpha(α), which is also know as the significance level. By choosing a significance level researcher can control the risk of making a Type I error.

↳ Note
Type I error matlab ⇒ Null Hypothesis actualy true tha, par humne galat reject kar dey

Type 2 error (false Negative)

↳ Type II error occur when based on the sample result, the Null Hypothesis is NOT rejected when it is in fact false.

This means that the researcher fail to detect a significant effect or relationship when one actually exists. The Probability of committing a type II error is denoted by β (Beta).

Note → Null Hypothesis galat tha, par tumne 'sahi' maan leya (fail to 'reject').

Truth about the Population

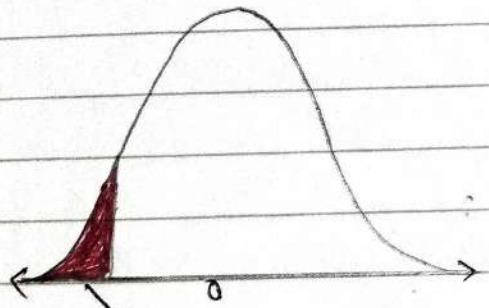
Decision based on sample	H_0 true	H_0 false
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type 2 error

One sided vs two sided test

One-sided (one tailed) test

↳ A one sided test is used when the researcher is interested in testing the effect in a specific direction. (either greater than or less than the value specified in the null hypothesis). The alternate hypothesis is a one-sided test contains an inequality (either ">" or "<").

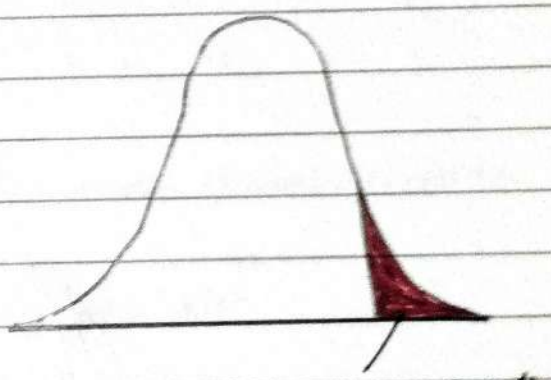
Ex → A researcher want to test whether a new medication increase the avg recovery rate compared to the existing medication.



Ha is more probable.

↳ left tail test

Ha : μ < value



Ha is more probable

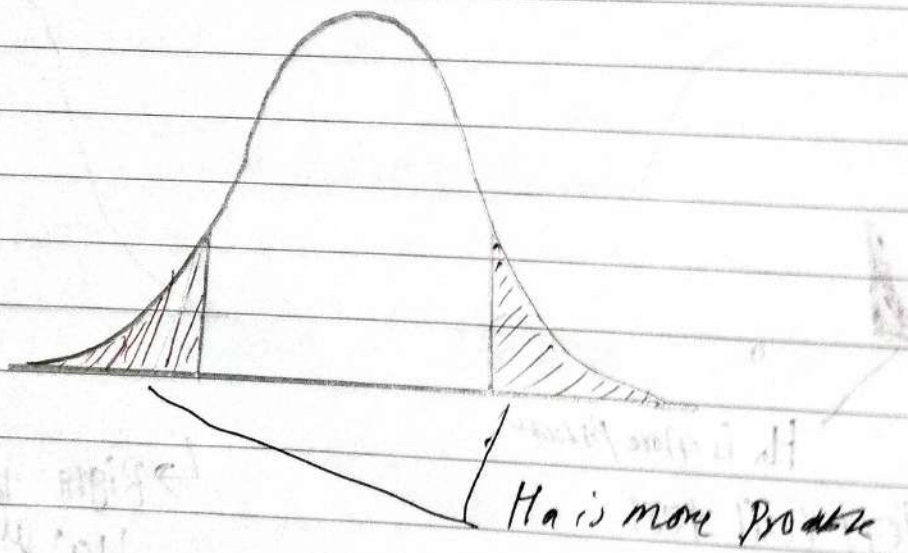
↳ Right tail test

Ha: μ > value.

Two-sided (two-tailed) test:

A two-sided test is used when the researcher is interested in testing the effect on both ~~side~~ direction (ie, whether the value specified in the null hypothesis is different, ~~either~~ greater or lesser). The alternate hypothesis is a two-sided test contains a "not equal to" sign (≠)

Ex ⇒ A researcher want to test whether a new medicine has diff avg recovery rate compared to the existing medication.



The main difference between them lies in the directionality of the alternate hypothesis and how the significant level is distributes in critical region.

P-value Approach

→ Pvalue is a measure of the strength of the evidence against the null hypothesis that is provided by our sample data.

Steps in P-value approach

① state hypothesis

② choose test type like z test, t-test, chi-square test, Anova

③ find the statistic for test;

or z test

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

④ find P-value using table:

• Z-table ya T-table se test statistic ke area (probability) dekho.

Us area ko depend karta hai test ka direction par

Right tailed: P-value = 1 − Z table value

left tailed: Pvalue = Z table value

2 tailed: Pvalue = 2 × (1 − table area) (agar right tail)

⑤ Compare with (significance level)

↳ P-value interpretation

$P < 0.01$ → Very strong evedience against H_0 → reject confidently

$0.01 \leq P \leq 0.05$ → moderate evidence against H_0 → Reject H_0

$0.05 \leq P \leq 0.10$ → weak evidence against H_0 → fail to reject H_0. but keep suspicion

$P \geq 0.10$ → No evidence against H_0 → fail to reject H_0

P-value in context of z-test

Suppose a company is evaluating the impact of new training program on the productivity of its employees. The company has data on the avg productivity of its employees before implementing the training program. The avg productivity was 50 units per day. After implementing the training program, the company measure the productivity of a random of sample of 30 employees. The sample has an average productivity of 53 units per day and population std is 4. The company wants to know if the new training program has significantly increased productivity.

for

$\mu = 50 \quad n = 30 \quad \bar{x} = 53$

$\sigma = 4 \quad \alpha = 0.05$

$H_0 : \mu = 50$

$H_a : \mu > 50$

$$\text{Z-Stat} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{53 - 50}{4 / \sqrt{30}} \approx 4.10$$

P-value

for $z \approx 4.11$ (right tailed)

P-value ≈ 0.00002 (basically 0)

$P < \alpha$, we can reject H_0