Date .....

* Introduction to Probabilities -

Probability is defined as the chance of happening or occurences of an event. Generally, the possibity of analyzing the occurence of any event with respect to previous data is called probability.

For example, if a fair coin is tossed, what is the chance it lands on the head? These types of questions are answered under probability.

- What is Probability Theory -

It uses the concept of random variables and probability distribution to find the outcome of any situation. Probability theory is an advanced branch of mathematics that deals with the odds and statistics of happening an event.

- Theoritical Probability -

It deals with assumptions in order to avoid unfeasible or expensive repition of experiments. The theoretical Probability for an Event A can be calculated as -

$$P(A) = \frac{\text{(Number of outcomes favourable to Event A)}}{\text{(Number of all possible outcomes)}}$$

- Experimental Probability -

It is found by performing a series of experim-ents and observing their outcomes. These random experiments are also known as trial. The experime-ntal probability for Event A can be calculated as -

$$P(E) = \frac{\text{(Number of times event A happend)}}{\text{(Total no. of trials)}}$$

Date .....

- Basics of Probability Theory -

- **Random Experiment -**
  Any event which can be repeated multiple times and its outcomes is not hampered by its repetition is called a random experiment. Tossing a coin, rolling a dice etc.

- **Sample Space -**
  The set of all possible outcomes for any random experiment is called sample space. For example, throwing dice results in six outcomes, which are 1, 2, 3, 4, 5 and 6. Thus, its sample space is (1, 2, 3, 4, 5, 6).

- **Event** - The outcome of any experiment is called an event.

Various types of events used in probability -

- **Independent Events** - The event whose outcomes are not affected by the outcomes of other future or past event. For example, tossing a coin.
- **Dependent Events** - The events whose outcomes are affected by the outcome of other events. For eg, Picking oranges from a bag without replacement.
- **Mutually Exclusive Events** - The event that cannot occur simultaneously. For eg, obtaining a head or a tail in tossing a coin, because both (h & t) can't obtained together.
- **Equally likely Events** - The events that have an equal chance or probability of happening are known as equally likely events. For eg, observing any face is rolling dice has an equal probability of 1/6.

Date .....

- Random Variable -

A variable that can assume the value of all possible outcomes of an experiment is called a random variable in Probability Theory. Random variables in probability theory are of two types which are discussed below,

Discrete Random Variable - Variables that can take countable values such as 0, 1, 2.... are called discrete random variable.

Continuous Random Variable - Variables that can take an infinite number of values in a given range is called continuous random variables.

- Permutation and Combination -

They are the way to represent a group of objects by selecting them in a set and forming subsets. It defines the various ways to arrange a certain group of data.

- Permutation - It relates to the act of arranging all the members of a set into some sequence or order. In other words, if the set is already ordered, then the rearranging of its elements is called the process of permuting.

- Combination - It is a way of selecting item from a collection, such that (unlike permutation) the order of selection does not matter. In smaller cases, it is possible to count the number of combination.

PERMUTATION - ARRANGEMENT

COMBINATION - SELECTING / CHOOSING

Date .....

- Common terms used in Permutation

Number of all permutation (P) of n things taking r at a time.

Eg- 3 alphabets (a, b, c) = n
Taking 2 at a time = r
arranging the alphabets.

Formula = $^nP_r = \frac{(n!)}{(n-r)!}$

$= \frac{3!}{(3-2)!} = 3! = 6$

- Common terms used in Combination

Number of all combination (C) of n things taking r at a time.

Eg- Same example and in this select the alphabets.

Formula = $^nC_r = \frac{n!}{r!(n-r)!}$

$\Rightarrow n = 3 \quad , r = 2$

$\Rightarrow \frac{3!}{2!(3-2)!} = \frac{3 \times 2!}{2!} = 3$

$\therefore$ 3 combinations are possible.

★ No. of all Permutation for n things taken all at a time = $^nP_n = n!$

★ No. of all Combination for n things taken all at a time = $^nC_n = 1$

Spiral

Date .....

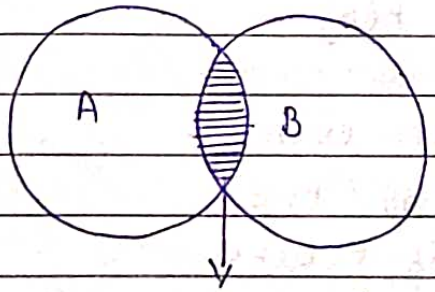
- D/B Permutation and combination -

Permutation	Combination
• Arranging peoples, digits, numbers, alphabets, letter and colors.	• Selection of menu, food, clothes, subject, team.
• Picking a team captain, pitcher and shortstop from a group.	• Picking three team members from a group.
• Picking two favourite colors, in order, from a color brochure.	• Picking two colors from a colour brochures.
• Picking first, second and third place winners.	• Picking three winners.

- UNION AND INTERSECTION -

↓ complete set

↓ common sample spaces b/w two sets.



Venn Diagram

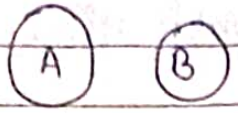
↓ Intersection

Notation for event A and event B intersection, $A \cap B$ or A and B

Date .....

• Cases in Intersection / Interaction -

① Null intersection



It is denoted as $\phi = \{\}$

- UNION - The union of set A and set B is equal to the set containing all the elements in A and B. This is represented as A∪B or A or B.

$$A \cup B = A + B$$

② If the event intersects then the $A \cup B =$

$$A \cup B = A + B - A \cap B$$

- Formula = $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

* Conditional Probability -

It is a probability of occurence of any event A, when event B in relation to A has already occured. This also means probability of event A depends on another event B.

Formula = $P(A|B) = \frac{P(A \cap B)}{P(B)}$

* When A and B are independent events -

$$P(A|B) = P(A) \cdot P(B) / P(B)$$

$P(A|B)$ is not defined if $P(B) = 0$

Cases - ① If A and B are disjoint

Then $A \cap B = \phi$

So, $P(A|B) = 0$

Date .....

When A and B are disjoint they cannot both occur at the same time. Thus, given that B has occured, the probability of A must be zero.

② B is a subset of A
Then $A \cap B = B$

③ A is a subset of B
Then $A \cap B = A$
So, $P(A|B) = P(A \cap B) / P(B)$
$\Rightarrow P(A|B) = P(A) / P(B)$

- Joint Probability - It is the probability of event A and event B happening, P(A and B).
It is the likelihood of the intersection of two or more events. The probability of the intersection of A and B is written as $P(A \cap B)$.

Joint Probability (JP) = $P(A) \times P(B)$

* Bayes Theorem -
It is also known as the Bayes Rule or Bayes Law. It is used to determine the conditional probability of event A when event B has already happened. The general statement of Baye's Theorem is "The conditional probability of an event A, given the occurrence of another event B, is equal to the product of the event of B, given A and the probability of A divided by the probability of event B". i.e.,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where, P(A) and P(B) = Probability of event A and B.
P(A|B) = Probability of event A when event B happens.

Date .....

$P(B|A)$ = vice versa of $P(A|B)$.

Terms Related to Bayes -
Conditional Probability
Joint Probability
Random Variable

D|B Conditional Probability and Bayes Theorem -

Baye's Theorem	Conditional Probability
- It is derived using the definition of conditional probability. It is used to find the reverse probability.	- It is the probability of event A when event B has already occured.
- Formula - $P(A\|B) = \frac{[P(B\|A)P(A)]}{P(B)}$	- Formula - $P(A\|B) = \frac{P(A \cap B)}{P(B)}$

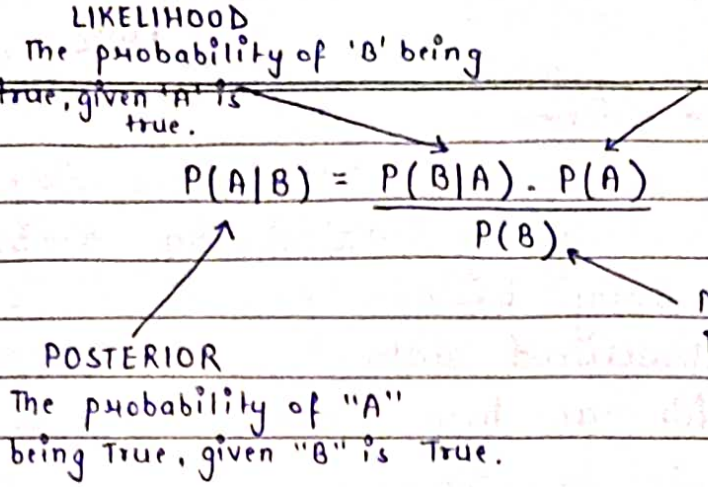
→ Basic terminology of Baye's theorem formula -

Hypothesis - The event A is called hypothesis.

Evidence / Data - The event B is called Evidence / Data.

And we read the $P(A|B)$ as to find the probability of hypothesis given that we have been given a data or we have observed some evidence.

Date .....



* STATISTICS -

It is a field of maths that deals with the collection, analysis, explanation and presentation of data. Statistical methods can be used to find solutions to problems which contains numerical data. These mathematical formulas or statistical method used to find solutions are known as quantitative models. Thus, we can define Statistics as a branch of mathematics which uses different quantitative models to produce a solution for a real-world problem containing numerical data. Statistics can be used to solve real-world complex problems using data. knowledge of Stastics can help a person to draw some very interesting inferences from data.

- Real world example of application of Statistics -

- Collection of data in data-surveys.
- Filling of missing values in data (Data cleaning)
- Analysis of Data
- Finding useful inferences from data
- Presentation of data using charts and tables (Boxplot, Histogram, etc).

Date .....

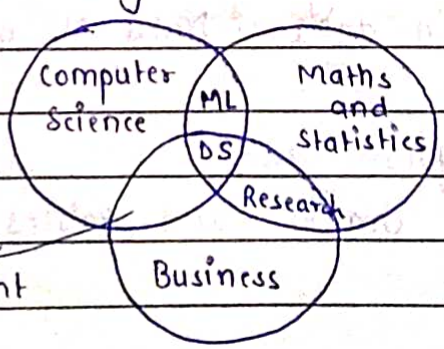
- Roles of Statistics in Data science -

Data is a term that we usually hear these days, enormous amount of data is being generated by people all around the world on a daily basis.

Data scientist use unstructured data to find some really interesting inferences which can help their companies in many ways.

- But finding useful info from the raw data isn't an easy job, the data that we gather require a lot of preprocessing. Only after preprocessing the data then the data can be used for visualizations and training models. Here, statistics comes into play, having a good knowledge of statistics can help a programmer in various steps of Data Analysis.
- It can be used in data pre-processing, for example seeing the unstructured data an analyst can determine whether to drop the rows which do contains any value or fill them with mean, mode etc.
- After data pre-processing, an analyst can find various inferences from data even without visualizing the data.
- Statistical visualizations like boxplot, histogram, line etc. can be very important to convey these inferences to the stakeholders.

And finally, when we train a ML model, Statistics can be of great help in evaluating our model, preprocessing the data before fitting in the model etc.



Date .....

Overall, stastistics is essential in D.S because it provides a framework for understanding and analyzing data. Without statistics, data scientists would not be able to make sense of the vast amounts of data that is generated in today's world, nor would they be able to use that data to make informed decisions.

Some of the key applications of statistics -

① Descriptive statistics - It is a set of statistical techniques that are used to summarize and describe the characteristics of a dataset. This includes measures of central tendency (such as mean, mode and median), measure of variability (such as range and standard deviation), and measures of distribution (such as histogram and frequency table).

② Inferential Statistics - This involves making inferences about a population based on a sample of data, using techniques such as hypothesis testing and confidence intervals.

Hypothesis testing involves making a statement about a population parameter (such as population mean) and checks whether the sample data supports or contradicts that statement.

Confidence Intervals are another inferential statistical technique used to estimate population parameter. For eg, a 95% confidence level for the population mean would be a range of values that is expected to contain the true population mean with 95% confidence.

Date .....

- Data and Their types -

Data is defined as a systematic record corresponding to a specific quantity. Basically, data can be summarized as a set of facts and figures which can be used to serve a specific usuage or purpose.

Types - ① Source - The origin or location of data, from where it is collected or obtained.
- Primary data - Firsthand data which is collected by researchers through surveys, experiment and interviews.
- Secondary data - obtained from existing source such as government reports, databases,

② Organization - The structure of arrangement of the data, which can be structured (organized in specific way, such as in table, database and spreadsheet) or unstructured (not organized in particular way, such as free text, images and audio).

③ Values - Different categories or types of data values that can be measured, such as numerical, categorical or textual.

- Qualitative Data - Qualitative data is used to represent som characteristics or attributes of the data. For instance, data on attributes such as honesty, loyalty, wisdom and creativity for a set of persons defined can be considered as qualitative data. Example - Attribute of peopl to a political system, Music and art, intelligence, Beauty of a person.

It can be further divided into two parts -

Spiral

Date .....

① Nominal data - It is used to label the variable without providing the numerical value to them. It can be both qualitative and quantitative in nature. The visualization of this data is done using pie charts. Example - Gender, Eye Color, Hair color, Marital status.

② Ordinary data - It is specific type of data that follows a natural order. The difference is the data values is not determined in the case of nominal data. For instance, ordinary data variable is mostly found in surveys, economics and finance operation. Example - Feedback is recorded in the form of rating 1-10. Economic status - poor, rich, medium, Letter grades - A, B, C etc.

• Quantitative data - It can be measured and is not just observed. The measurement of data is numerically recorded and represented. Calculations and interpretations can then be performed on the obtained results. Numerical data is indicated by quantitative data. For instance, data can be recorded about how many user found a product satisfactory in terms of the collected rating, and therefore, an overall product review can be generated. Example - Daily Temp, Price, Weight, Income etc. It can be further divided into 2 - Discrete and continuous

① Discrete data refers to the data values which can only attain certain specific values. Discrete data can be represented using bar charts. For instance, ratings of a product made by the users can only be in discrete no. Eg - The no. of students in a class, number of chips in a bag, The no. of stars in the sky.

② Continuous data - height and weight of a student, daily temp. recording of a place, wind speed measurement.

Date .....

④ Variables - The characteristics or attributes of the data that can be measured or observed, such as age, height, weight, gender etc.

- Univariate Data - It refers to an analysis or data that involves a single variable. Univariate analysis is useful for understanding the behavior of a single variable and can help identify patterns or trends in the data.
- Bivariate Data - It refers to an analysis or data that involves two variables. It is used to explore the relat-ionship b/w two variables, such as their correlation or causation. It can be done using different statistical techniques, such as correlation analysis, regression analysis.

* Descriptive Statistics -

- Frequency distribution -

It is a tabular summary of data showing the frequency (or number) of items in each of several non-overlapping classes.

It is of two types - ① Un-grouped - All distinct values of variable are mentioned and their frequencies are counted.

② Grouped - Values are divided b/w different intervals and then their frequencies are counted.

Eg- import pandas as pd
import numpy as np
import seaborn as sns # used for statistical graphic making.

Date .....

```
from matplotlib import pyplot as plt
    df = sns.load_dataset(" ")
        df.head()

df[" "].value_counts().plot(kind = "bar")
```

* Measure of Central Tendency -

This measure is an important way to summarize the dataset with one representative value. This measures provides a rough picture of where data points are centered The commonly used measures of central tendency are -

- Mean
- Median
- Mode

• Mean - "Average" value is termed as the mean of the dataset. It is very easy to calculate the mean.

Steps to calculate mean -

① Count the no. of data value. Let it be n.

② Add all the data values. Let the sum be s.

③ Mean = Sum of all data values (s) / Total no. of data values

$$\text{Mean } (\bar{x}) = \frac{\Sigma x}{n}$$

```
# Sample data
    arr = [5, 6, 11]
# Mean
    mean = np.mean(arr)
        print("Mean =", mean)
```

O/P - Mean = 7.3333

Date .....

- Median - The middle value of the sorted dataset is called median, Consider a dataset comprising 'n' elements.

Steps to calculate median -

① The dataset is arranged in either increasing or descending order.

② If the data set has an odd no. of data values (n=odd) then the middlemost value of the sorted dataset is computed as the median. In other words, the data at (n+1)/2 place is the median of the dataset.

③ If the dataset has an even no. of data values (n = even), the average of two middle values is computed as the median. i.e. the mean of (n/2) and (n/2 + 1)th is the median of the dataset.

Odd	Even
$\frac{n+1}{2}$	$\frac{n}{2}, \frac{n}{2}+1$

- Mode - The most frequently occuring value in the dataset

Steps to calculate mode -

① Use tally marks to identify how many times each data value occurs in the dataset.

② The data value with maximum tally is the mode of the dataset.

* Measure of Dispersion -

They are statistical values that help in understanding how much the data in a dataset varies or is spread out from its central tendancy. Dispersion indicates how much the values in a dataset deviate from the mean or median.

Spiral

Date .....

In other words, it provides a measure of the diversity or variability of the dataset. The most common measures of dispersion are the range, variance and standard deviation.

- Range - It is the simplest measure of dispersion and it represents the difference between the highest & lowest value in a dataset. It provides a rough idea of how far the data spread.

  Range = Largest data value - Smallest data value

- Variance - It is the average of the squared deviation of each data point from the mean. It measures how far the data points are from the mean. Variance is calculated by squaring the difference b/w each data point and the mean, adding all of these squared difference, and then dividing the sum by the total no. of data points in the dataset.

$$\text{Variance} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

x = data points
n = total no. of data points in the dataset
$\bar{x}$ = mean of the dataset. (15-3)

Code -

```
# sample data
arr = [1, 2, 3, 4, 5]
arr2 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
# variance
print ("Var = ", (statistics.variance(arr)))
print ("Var =", (statistics.variance(arr2)))
```

O/P - Var = 2.5

Date .....

- **Standard Deviation** - It is the square root of the variance. It is widely used measure of dispersion as it is easy to interpret and has desirable mathematical properties. It indicates how much the data points deviate from the mean in term of standard deviation. A small S.D indicates the data points are tightly clustered around the mean, while a large standard deviation indicates that the data points are spread out from the mean.

$$S.D = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- **Coefficient of Variation** - It is the measure of dispersion that expresses the standard deviation as a percentage of the mean. It is used to compare the variability of datasets with different means.

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\bar{x}} \times 100$$

where,

$\bar{x}$ = mean of the dataset.

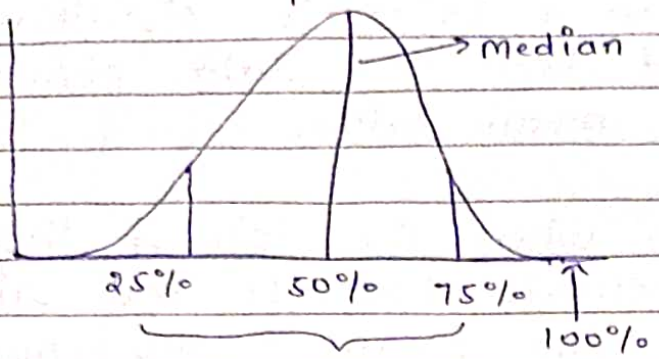
- **Quantile** - (Interquartile Range)

Quantile determines how many values in a distribution are crossing a threshold i.e., how many values are above and below a certain limit. The interquartile range is a measure of dispersion that is based on the quartile of the dataset.

Spiral

Date .....

The quartiles divides the whole dataset into four equal parts, with each part representing 25% of the data.



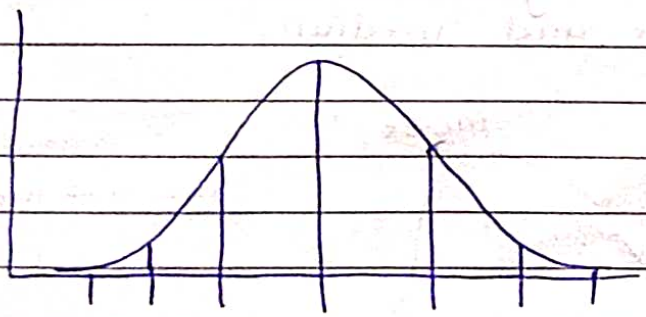
Interquartile Range = $Q_3 - Q_1$

* Measure of shape -

When analyzing data, it is essential to not only understand the central tendency and variability of the data but also its shape. The shape of the distribution tells us about the pattern of the data and how the values are distributed around the central tendency.

- Normal Distribution -

When we plot a dataset such as a histogram, the shape of that charted plot is what we calls its distribution. The most commonly observed shape of continuous value is the bell curve, also called the Gaussian or normal distribution.

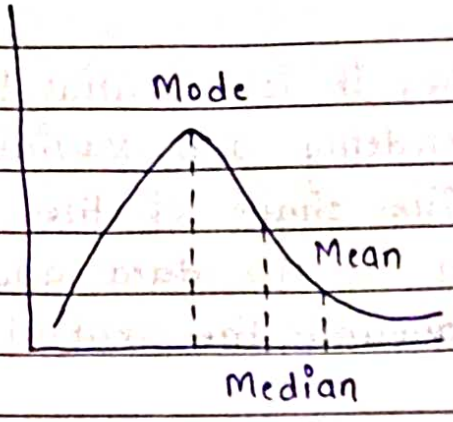


Date .....

- **Skewness -**

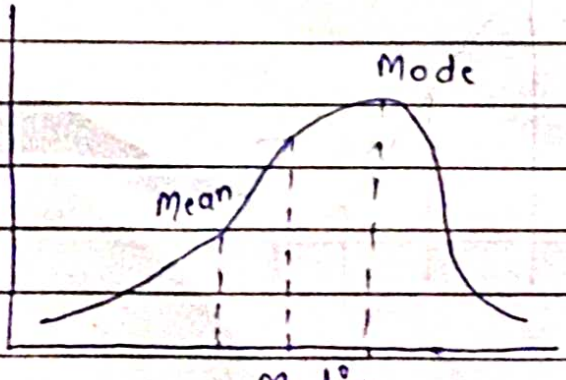
It is a statistical measure that describe the degree of asymmetry in a dataset's distribution. It is used to understand how the data points are distributed around the mean value.

① **Positive Skewness** - when the tail of the distribution extends towards the right-hand side of the curve. The mean of the positively skewed distribution is greater than the mode and median.



mode < median < mean

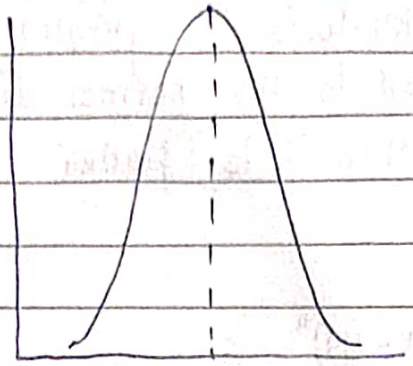
② **Negative Skewness** - A distribution is said to be negatively skewed when the tail of the distribution extends towards the left-hand side of the curve. The mean of a negatively skewed distribution is less than the mode and median.



mean < median < mode

Date .....

• Normal Skewness - The distribution that is perfectly symmetrical. The mean, mode and median are equal. This means equal number of data points are on the both hand side.



Symmetrical data mean = median = mode

→ Skewness is a statistical number thats tells us if a distribution is symmetric or not.

A distribution is symmetric if the right side of the distribution is similar to the left side of distribution.

- Pearson's Coefficient of Skewness -

$$S_K = \frac{\text{Mean} - \text{Mode}}{\sigma \text{ (Standard deviation)}}$$

If mode is not well defined, we use the formula

$$S_K = \frac{3(\text{Mean} - \text{Mode})}{\sigma}$$

Range of skewness = -1 or 1

- To increase the skewness use square and to reduce the skewness use square root.

Date .....

- Kurtosis - It is a statistical number that tells us if a distribution is taller or shorter than a normal distribution.

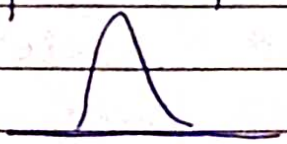
If a distribution is similar to the normal distribution, the kurtosis value is 0. If kurtosis is greater than 0, then it has a higher peak compared to the normal distribution. If kurtosis is less than 0, then it is flatter than a normal distribution.

$$\text{Kurtosis} = \frac{E[(X-\mu)]^4}{\sigma^4} - 3$$

Code ⟶ df. kurtosis ()

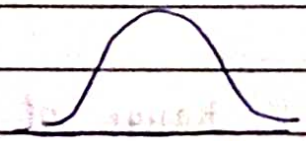
There are three types of distribution:

- Leptokurtic - Sharply peaked with fat tails, and less variable.



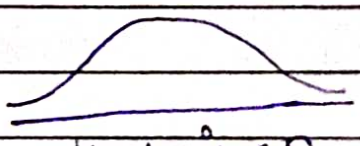
Kurtosis > 0

- Mesokurtic - Medium peaked



Kurtosis = 0

- Platykurtic - Flattest peak and highly dispersed



Kurtosis < 0