Probability (Tutorial 01)

★ **Why learn Probability ?**

Understanding probability gives you the ability to predict how future events may turn out.

★ **What is Probability ?** (Normally, denoted as P.)

A probability provides a quantitative description of the chances or likelihoods associated with various outcomes.
Probability is a measure of how likely an event or outcome is.
Probability is simply how likely something is to happen.

★ **Experiments (trial) and Outcomes**

Experiment – An experiment is any process that generates well defined outcomes.

- ➔ Tossing a coin
- ➔ Rolling a die
- ➔ Measuring height or weight
- ➔ Selecting a card from a card pack
- ➔ Number of vehicles passing through a certain place

Outcome – possible result of an experiment

Eg: If you toss 2 coins the four possible outcomes are
HH,HT,TH,TT

If you toss 1 coin the Possible outcomes are Head or Tail.
HH,TT

★ **Random Experiment**

Random experiment is an experiment or a process for which the outcome cannot be predicted with certainty.

In random experiments, events can be of the following types,

- ➔ **Certain events** (the event is definitely going to happen)
  Ex:- If it is Thursday, the probability that tomorrow is Friday

- ➔ **likely event** (the event will probably happen, but not definitely)
  Ex:- Probability of getting head if a coin is tossed

- ➔ **Unlikely event** (the event will probably not happen, but it might)
  Ex:- your hair would grow 5 inches longer by tomorrow

- ➔ **Impossible event** (the event is definitely not going to happen)
  Ex: the probability that tomorrow is Friday if today is Monday

- ➔ **Simple event** (an event corresponds to a single possible outcome of the experiment)
  Ex: throwing a die S={1, 2, 3, 4, 5, 6}
  The event of getting 5 on the uppermost face

★ Sample space

Sample space is the set of all possible outcomes

Eg:
Toss a fair coin
S={ H, T }

Rolling a Die
S={ 1,2,3,4,5,6}

★ Event

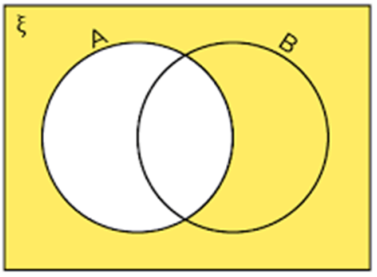
An event is an outcome of an experiment

➔ Experiment: Tossing a coin
Sample space = {H,T}
Event: getting head {H} ,
getting tail {T}

➔ Experiment: Toss a die
Sample space = {1,2,3,4,5,6}
Event A: observe an odd number {1,3,5}
Event B: observe a number greater than 2 {3,4,5,6}

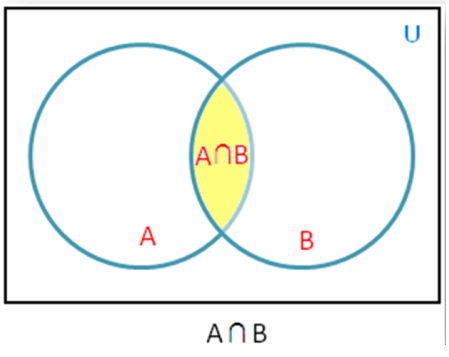
Complement

The complement of an event A, w.r.t. S is the subset of all elements of S that are **not in A** and are denoted by A**'** **or** A^c**.**



Intersection

The intersection of two events A and B denoted by the symbol **A ∩ B** is the event containing all the elements that are **common to A and B.**



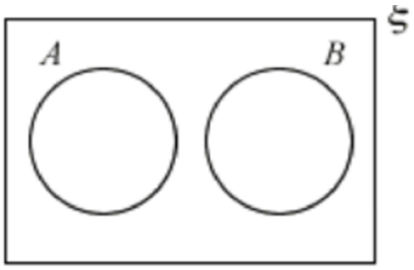
Mutually Exclusive Events

If two events cannot occur at the same time they are called mutually exclusive events

Examples:

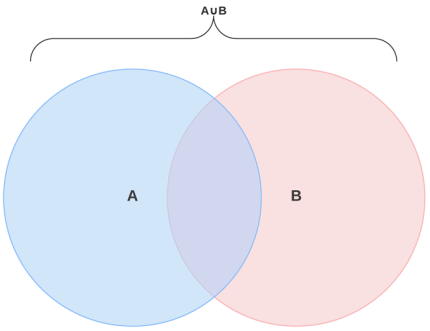
- ➔ Turning left and turning right are Mutually Exclusive (you can't do both at the same time)
- ➔ When tossing a coin, the event of getting head and tail are mutually exclusive.

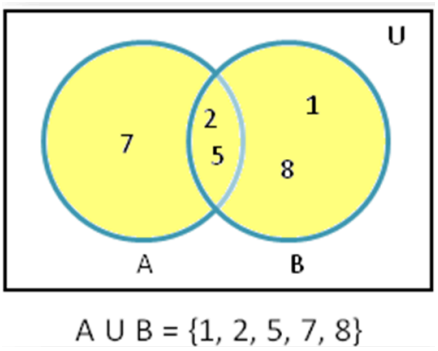
` For mutually exclusive events $P(A \cap B) = 0$



Union

The union of two events A and B, denoted by the symbol A ∪ B is the event containing **all the elements that belong to A or B or both.**



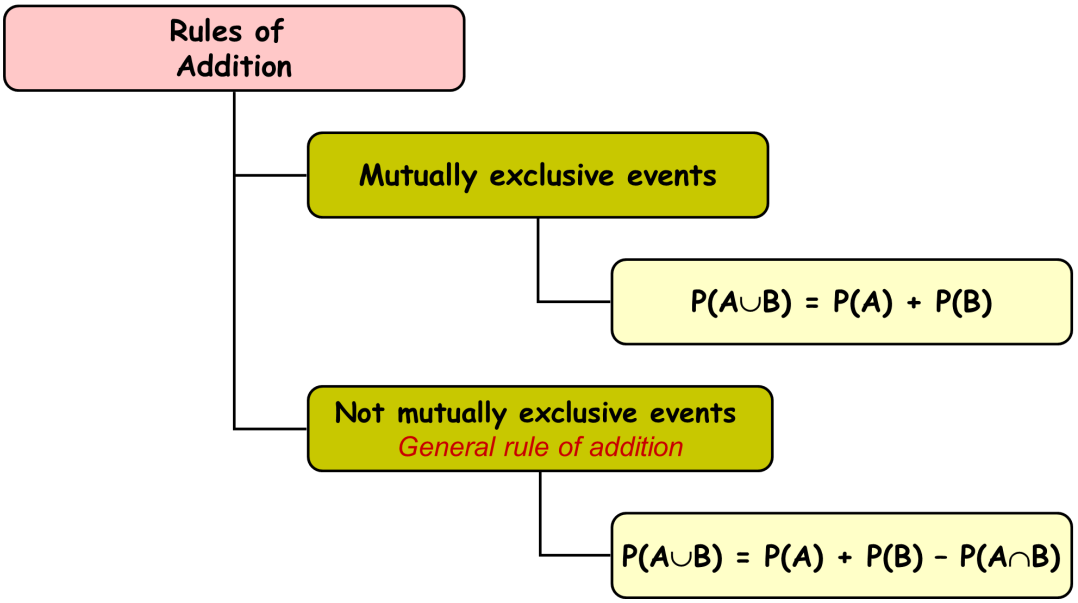


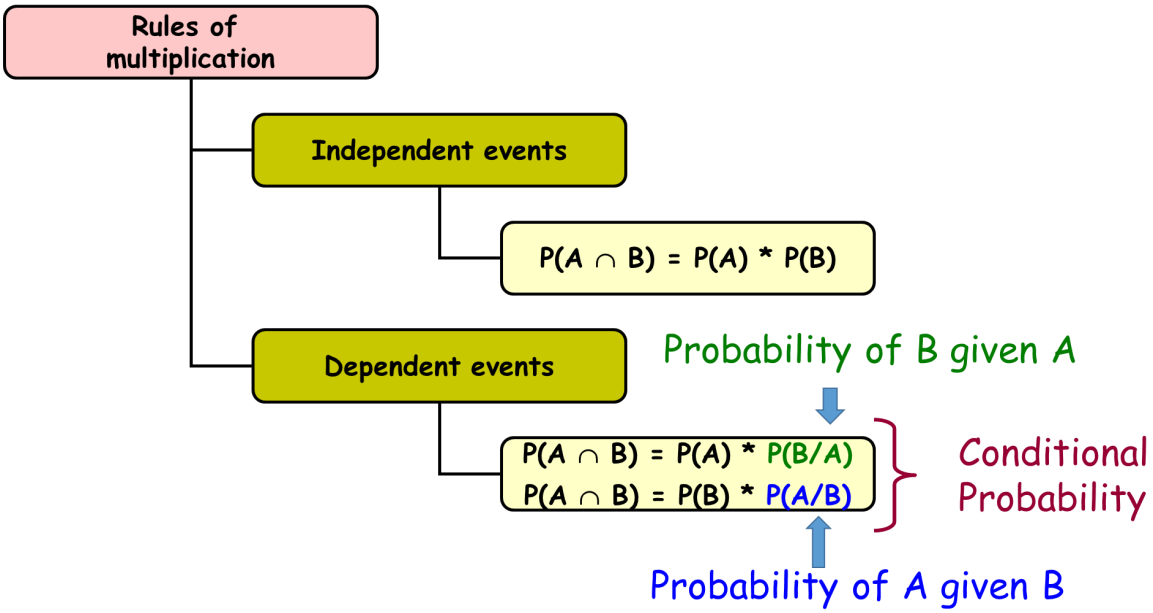
A U B = {1, 2, 5, 7, 8}

★ **Probability of an Event**

n(A) – number of elements in the set of the event A
n(S) – number of elements in the set of the sample space
P(A) – probability of event A

P(A) = $\frac{n(A)}{n(S)}$





★ Mutually Exclusive Events

If two events cannot occur at the same time they are called mutually exclusive events

Eg: When tossing a coin, the event of getting head and tail are mutually exclusive.

For mutually exclusive events P (A ∩ B) = 0

★ Independent Events

If the occurrence of an event A does not affect the occurrence of event B then A and B are independent events

Eg: simultaneously tossing two coins

For independent events

$$P(A \cap B) = P(A) \times P(B)$$

★ Conditional Probability

Conditional probability is a measure of the probability of an event occurring, given that another event has already occurred.

The conditional probability of event B given the occurrence of event A is

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

★ Rules of Probability

➔ Additive Rule

If A and B are any two events, then

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

If A, B, and C are any two events, then

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

If A and B are mutually Exclusive events, then

$$P(A \cup B) = P(A) + P(B) \quad ; \quad Since \quad P(A \cap B) = 0$$

➔ Law of Complement

If the probability of an event A is denoted by P(A), then

$$P(A') = 1 - P(A)$$

➔ Rule of Multiplication

$$P(A \cap B) = P(A) \times P(B)$$

Example: 01

Suppose you know that the probability of getting the flu this winter is 0.43. What is the probability that you will not get the flu?

$P(A) = 0.43$

$P(A') = 1 - 0.43$

$= 0.57$

Example: 02

In a group of 40 people, 10 are healthy and every person of the remaining 30 has either high blood pressure, a high level of cholesterol, or both. If 15 have high blood pressure and 25 have a high level of cholesterol,

I. How many people have high blood pressure and a high level of cholesterol ?

$$n(H.B.P \cup H.L.C) = n(H.B.P) + n(H.L.C) - n(H.B.P \cap H.L.C)$$
$$30 = 15 + 25 - n(H.B.P \cap H.L.C)$$
$$10 = n(H.B.P \cap H.L.C)$$

II. If a person is selected randomly from this group, what is the probability that he/she;

a) has high blood pressure?

$$P(H.B.P) = \frac{n(H.B.P)}{n(total)}$$
$$= 15/40$$
$$= 0.375$$

b) has a high level of cholesterol?

$$P(H.L.C) = \frac{n(H.L.C)}{n(total)}$$
$$= 25/40$$
$$= 0.625$$

c) has high blood pressure and high levels of cholesterol ?

$$P(H.B.P \text{ and } H.L.C) = \frac{n(H.B.P \text{ and } H.L.C)}{n(total)}$$
$$= 10/40$$
$$= 0.25$$

d) has either high blood pressure or a high level of cholesterol ?

$$P(H.B.P \text{ or } H.L.C) = \frac{n(H.B.P \text{ or } H.L.C)}{n(total)}$$
$$= 30/40$$
$$= 0.75$$

III. Use part II to verify the addition law of probability

$$P(H.B.P \cup H.L.C) = P(H.B.P) + P(H.L.C) - P(H.B.P \cap H.L.C)$$
$$= 15/40 + 25/40 + 10/40$$
$$= 30/40$$

Example: 03

Employees of a company are classified as follows:
Supervisors 170, Managers 20, Secretaries 10

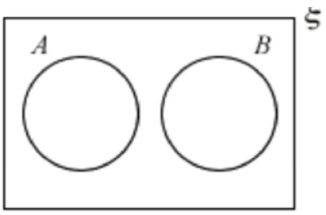
I. What is the probability that the first person selected either a supervisor or a manager

$$P(A \cup B) = \frac{P(A)}{N} + \frac{P(B)}{N}$$
$$= 170/200 + 20/200$$
$$= 190/200$$
$$= 0.95$$

This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

II. What is the probability that the first person is not a manager. Draw a Venn diagram for 1 and 2.

$$1 - P(A \cup B) = 1 - 0.95$$
$$= 0.05$$



III. Are the events complementary or mutually exclusive or both

event are mutually exclusive because one persian can't act two character at the same time

Example: 04

A survey of MDA students reveals that
35% of them read Sunday Times,
20% reads Sunday Observer and
40% reads Sunday Island newspapers.
10% reads both Sunday Times and Sunday Island newspapers.

I. What is the probability that a particular student read either Sunday Times or Sunday Island?

st = 0.35
so = 0.2
si = 0.4
st and si = 0.1

$P(ST \cup SI) = P(SI) + P(ST) - P(ST \cap SI)$
$= 0.4 + 0.35 - 0.1$
$= 0.65$

II. What is the above 10% called ?

Joint probability of ST and SI

III. Are these events mutually exclusive ?

No, these events are not mutually exclusive , because a person can read two newspapers or three newspapers at the same time.

This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

Example: 05

A survey indicated that out of 250 persons interviewed, 107 read the Dinamina newspaper and around 83 read Daily News, and around 52 read both newspapers.

I. Compute the probability that a randomly chosen person will read at least one of the newspapers Dinamina or Daily News.

$N = 250$
$n(D) = 107,\ P(D) = 107/250$
$n(D.N) = 83,\ P(D.N) = 83/250$
$n(D\&D.N) = 52,\ P(D\&D.N) = 52/250$

$$P(D \cup DN) = P(D) + P(D.N) - P(D \cap D.N)$$
$$= 107/250 + 83/250 + 52/250$$
$$= 138/250$$
$$= 0.552$$

II. What is the probability that a randomly chosen person does not read any of the newspapers Dinamina or Daily News?

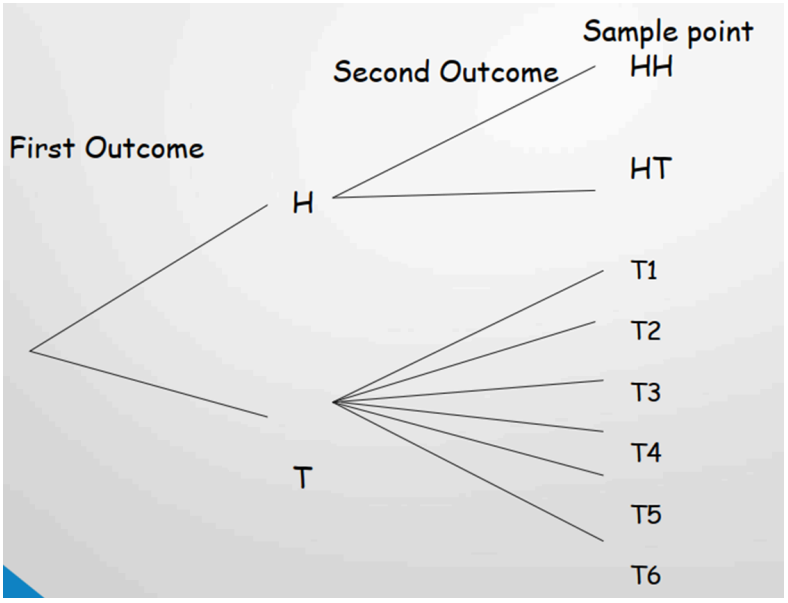
$$P(D \cup D.N)' = 1 - P(D \cup D.N)$$
$$= 1 - 0.552$$
$$= 0.448$$

Example: 06

An experiment consists of flipping a coin and then flipping it a second time if a head occurs. If a tail occurs on the first flip, then a die is tossed once.

To list the elements of the sample space providing the most information, we may construct the tree diagram.

S = { HH, HT, T1, T2, T3, T4, T5, T6 }



This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

Example: 07

Suppose you own two kinds of stocks, A and B. The probability that A stock will increase in value next year is 0.5 and the probability that B stock will increase in value next year is 0.7. Assume the two stocks are independent. What is the probability that both stocks will increase in value next year ?

$P(A) = 0.5$
$P(B) = 0.7$

$P(A \cap B) = P(A) \times P(B)$
$= 0.5 \times 0.7$
$= 0.35$

Random Variables and Probability Distributions (Tutorial 02)

★ Statistical Experiment

- ➔ Statistical experiment can have more than one possible outcome
- ➔ Each possible outcome can be specified in advance
- ➔ The outcome of the experiment depends on chance

★ Random Variable

A random variable is a numerical quantity that is generated by a random experiment.

When the value of a variable is the outcome of a statistical experiment that variable is a random variable

Consider the weights of 6-year-old children. Though they are of the same age, but their weights are different.

Therefore if we let X denote the weight of a 6-year-old child, X is a variable that takes different values. For a given instance those values are not certain. In other words, there are variations among observations due to unpredictable reasons.Such variation is called chance variation or random variation.

Notations
X – random variable X
P (X = x) The probability that the random variable X is equal to a particular value x
Eg: P (X=1) the probability that the random variable X is equal to 1

Table 4.1 Four Random Variables

Experiment	Number *X*	Possible Values of *X*
Roll two fair dice	Sum of the number of dots on the top faces	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Flip a fair coin repeatedly	Number of tosses until the coin lands heads	1, 2, 3,4, …
Measure the voltage at an electrical outlet	Voltage measured	$118 \le x \le 122$
Operate a light bulb until it burns out	Time until the bulb burns out	$0 \le x < \infty$

Types of random variables

➔ **Discrete random variables** – take only integer values

A random variable X is said to be discrete if it can assume only a finite or countably infinite number of distinct values.

For ex:
The sum of the number of dots on the top faces (x=2, 3, 4, 5, 6, 7, 8, 9, 10, 11,12)
The number of tosses until the coin lands head (x=1, 2, 3,4, …)

➔ **Continuous random variables** – take any value within a range of values

A random variable X that can take any value in a given interval is called a continuous random variable.

For ex:
Height of a person (x=100 cm, 100.52cm)
Amount of rainfall (x=120mm, 152.36mm)
Lifetime of a battery (x= 2.5 yrs, 3 yrs……)

★ **Probability Distributions**

Probability distribution is a table, an equation or a graph that links each outcome of a statistical experiment with its probability of occurrence.

A probability distribution is a table, equation, or graph that shows the chance of each outcome in a statistical experiment.

1. **Discrete Probability Distributions**

Deals with variables that take on a countable number of distinct values.

Examples:

- ➔ **Bernoulli Distribution:** Models a single binary outcome (e.g., success/failure).
- ➔ **Binomial Distribution:** Describes the number of successes in a fixed number of independent Bernoulli trials.
- ➔ **Poisson Distribution:** Represents the number of events in a fixed interval of time or space, assuming events occur independently.

2. **Continuous Probability Distributions**

Deals with variables that can take on an infinite number of values within a range.

Examples:

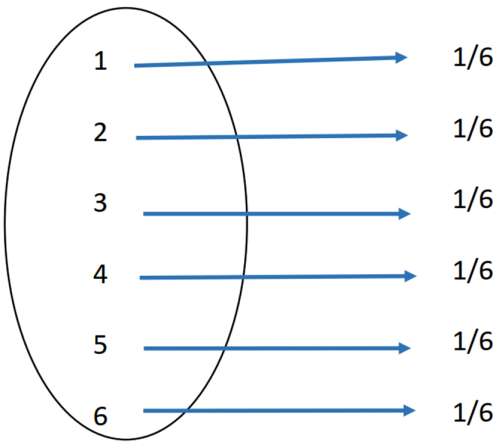
- ➔ **Normal Distribution (Gaussian):** Bell-shaped curve; most data falls near the mean.
- ➔ **Uniform Distribution:** All outcomes in a range are equally likely.
- ➔ **Exponential Distribution:** Describes the time between events in a Poisson process.
- ➔ **Gamma Distribution:** Generalization of the exponential distribution

1. Discrete Probability Distributions

Deals with variables that take on a countable number of distinct values.
The probability distribution of a discrete random variable can always be represented by a table.

Example 01

Consider rolling a fair die. Let X denotes the number facing up.



The probability distribution can be represented as follows:

x	1	2	3	4	5	6
P (X = x)	1/6	1/6	1/6	1/6	1/6	1/6

As each sample point is equally likely, the probability mass function of the random variable X can be written as:

$$P_X(x) = P(X = x) = \begin{cases} 1/6 & ; x = 1,2,3,4,5,6 \\ 0 & ; otherwise \end{cases}$$

Note that :

➔ The probability that X can take a specific value is p(x)
P (X=x) = p(x)

➔ p(x) is non negative for all real x
0 ≤ p(x) ≤ 1

➔ The sum of p(x) over all possible values of x is 1
∑ p(x) = 1

$P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)+P(X=6)$ =
1/6+1/6+1/6+1/6+1/6 +1/6 = 1

This is an identity that is always true.
P(□)=1 (the total probability is always equal to 1)

Example 02

Let X be a discrete random variable with a probability distribution given by:

x	0	1	2	3	4	5	6	7
P (X = x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2 + k$

a) Find the value of k.
Total probability always 1,

∑ p (X=x) = 1

$0 + k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$
$9k + 10\ k^2 = 1$
$10k^2 + 9k - 1 = 0$
$10k^2 + 10k - 1k - 1 = 0$
$10k\ (1+k) - 1\ (1+k) = 0$
$(10k - 1)\ (1+k) = 0$

(10k -1) = 0 or (1+k) = 0

10k = 1

k = $\frac{1}{10}$ k = -1 (not get the negative value)

b) Find each of the following probabilities:

i. $P\ (X \leq 6)$

$P\ (X \leq 6)$ = p(x=0) + p(x=1) + p(x=2) ….. + p(x=6)

= 0 + k + 2k + 2k + 3k + k^2 + $2k^2$

= 8k + $3k^2$

= 8 x $(\frac{1}{10})$ + 3 x $(\frac{1}{10})$

= 0.8 + $(\frac{3}{10})^2$

= 0.8 + 0.03

= 0.83

ii. $P\ (3 < X \leq 6)$

$P\ (3 < X \leq 6)$ = p(x=4) + p(x=5) + p(x=6)

= 3k + k^2 + $2k^2$

= 3k + $3k^2$

= 3 x $(\frac{1}{10})$ + 3 x $(\frac{1}{10})^2$

= 0.3 + 0.03

= 0.33

iii. $P\ (3 < X \mid X \leq 6)$

$X > 3$ = 4,5,6

$X \leq 6$ = 0,1,2,3,4,5,6

$3 < X \cap X \leq 6$ = 4,5,6

$P\ (3 < X \mid X \leq 6) = \frac{P\ (3 < X \cap X \leq 6)}{X \leq 6}$

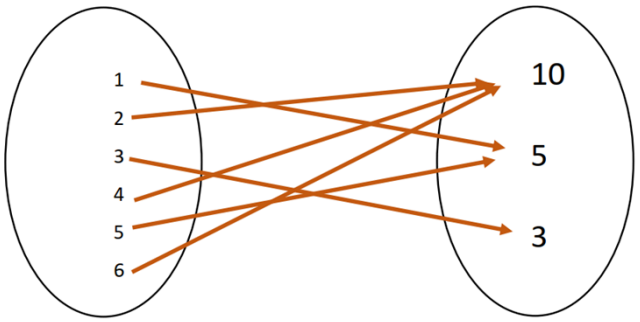
$= \frac{p(x=4) + p(x=5) + p(x=6)}{p(x=0) + p(x=1).....+ p(x=6)}$

= 0.33 / 0.83

= 0.397

Example 03

A game is played as follows, A die is rolled. If an even number turns up, the score of the game is regarded as 10. If the number turning up is 1, the score of the game is regarded as 5. O/W, the score of the game is regarded as the number facing up. Let X be the score of the game and find p.m.f. for X.



Sample space Induced sample space

x	3	5	10
P (X = x)	⅙	⅓	½

P (x = 3) = ⅙

P (x = 5) = ⅙ + ⅙
= ⅓

P (x = 10) = ⅙ + ⅙ +⅙
= ⅓

Example 04

A fair coin is tossed twice. Let X be the number of heads that are observed.

a) Construct the probability distribution of X.

S = { HH, HT ,TH, TT }

x (number of heads) = 2, 1, 1, 0

x	0	1	2
P (X = x)	¼	½	¼

b) Find the probability that at least one head is observed.

P (x ≥ 1) = P (x = 1) + P(x = 2)
= ½ + ¼
= ¾

The mean of a discrete probability distribution

- ➔ It is also known as expected value
- ➔ If the experiment is repeated many times the average value of the random variable is defined as the expected value

$$E(x) = \sum x \cdot p(x)$$

Example 05

Let X denote the number of cars sold by a particular company in a month. Suppose X has the probability distribution given below.The company is interested in finding out the expected number of cars sold in a month.

x	0	1	2	3	4
P (X = x)	0.15	0.30	0.35	0.15	0.05

E(x) = (0 x 0.15) + (1 x 0.30) + (2 x 0.35) + (3 x 0.15) + (4 x 0.05)
= 1.65
= Then, the expected number of cars sold in a month is 1.65 ≈ 2

Example 06

Let X be the number of defective items of a certain product in a batch of 100. From past experience, a manufacturer expects the probability distribution of X to be as given below.

x	0	1	2	3
P (X = x)	0.61	0.23	0.11	0.05

a) Compute the probability that there will be at most one defective item in a batch of 100 items.

$$P(x \leq 1) = P(x = 0) + P(x = 1)$$
$$= 0.61 + 0.23$$
$$= 0.84$$

b) Compute the probability that there will be at least two defective items in a batch of 100 items.

$$P(x \geq 2) = P(x = 2) + P(x = 3)$$
$$= 0.11 + 0.05$$
$$= 0.16$$

Or

$$P(x \geq 2) = 1 - P(x \leq 1)$$
$$= 1 - [P(x = 0) + P(x = 1)]$$
$$= 1 - [0.61 + 0.23]$$
$$= 0.16$$

c) What is the expected number of defective items in a batch of 100 items?

$E(x) = \sum x \cdot p(X = x)$
$= (0 \times 0.61) + (1 \times 0.23) + (2 \times 0.11) + (3 \times 0.05)$
$= 0 + 0.23 + 0.22 + 0.15$
$= 0.6$

Thus, the expected number of defective items is at most 1

The variance of a discrete probability distribution

$V(x) = E(x^2) - [E(x)]^2$ where,

$E(x^2) = \sum x^2 p(x)$

Example 07

Let X be a discrete random variable with the following probability distribution.find the variance and expected value.

x	0	1	2	3	4
P (X = x)	0.25	0.35	0.05	0.05	0.03

$E(x) = \sum x \cdot p(X=x)$
$= (0 \times 0.25) + (1 \times 0.35) + (2 \times 0.05) + (3 \times 0.05) + (4 \times 0.03)$
$= 0.72$

$E(x^2) = \sum x^2 \cdot p(X=x)$
$= (0^2 \times 0.25) + (1^2 \times 0.35) + (2^2 \times 0.05) + (3^2 \times 0.05) + (4^2 \times 0.03)$
$= 1.48$

$V(x) = E(x^2) - [E(x)]^2$
$= 1.48 - (0.72)^2$
$= 0.9616$

Example 08

Let X be a discrete random variable with probability mass function $P(X=x)$ given by

$$P(X = x) = \begin{cases} kx & ; x = 2,4,6 \\ k(x-2) & ; x = 8 \\ 0 & ; otherwise \end{cases}$$

a) Show that $k = \frac{1}{18}$

$$\sum p\,(X = x) = 1$$
$$(k \times 2) + (k \times 4) + + (k \times 6) + k(8 - 2) = 1$$
$$2k + 4k + 6k + 6k = 1$$
$$18k = 1$$
$$k = \frac{1}{18}$$

b) Construct the probability distribution of X.

x	2	4	6	8
P (X = x)	kx	kx	kx	k(x-2)
	$\frac{x}{18}$	$\frac{x}{18}$	$\frac{x}{18}$	$\frac{1}{18}$ (x-2)
	$\frac{2}{18}$	$\frac{4}{18}$	$\frac{6}{18}$	$\frac{6}{18}$
	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{1}{3}$

c) Calculate the expected value of X.

$E(x) = \sum x \,.\, p\,(X=x)$
$= (2 \times kx) + (4 \times kx) + (6 \times kx) + 8k\,(x - 2)$
$= (2 \times k \times 2) + (4 \times k \times 4) + (6 \times k \times 6) + 8k\,(8 - 2)$
$= 4k + 16k + 36k + 48k$
$= 104k$
$= 104 \times \frac{1}{18}$
$= 5.778$

d) Calculate the variance of X.

$E(x^2) = \sum x^2 \,.\, p(X=x)$
$= (2^2 \times kx) + (4^2 \times kx) + (6^2 \times kx) + 8^2\,(k(x - 2))$
$= 4k^2 + (16 \times k \times 4) + (36 \times k \times 6) + (64 \times k \times 6)$
$= 37.333$

$V(x) = E(x^2) - [E(x)]^2$
$= 37.333 - (5.778)^2$
$= 3.94$

This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

★ Binomial Distribution (Tutorial 03)

The probability distribution of a binomial random variable is called a binomial distribution.

An experiment consisting of n independent Bernoulli trials is called a Binomial experiment.

Binomial Experiment

It is a statistical experiment with following properties

- ➔ A fixed number of trials (n)
- ➔ Each trial should be success or failure
- ➔ The trials are independent
- ➔ The probability of success (p) at each trial is a constant

X ~ Bin (n,p)

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad ; for\ x = 0,1,2, \dots\dots\dots\dots n$$

$$= {}^{n}_{x}C\ P^{x}\ (1 - P)^{n - x} \qquad , \qquad {}^{n}_{x}C = \frac{n!}{x!\ (n-x)!}$$

Example 01

A coin is tossed 10 times. Find the probability of getting exactly 3 heads ?

n = 10 (number of trials),
x = 3 (number of successes, i.e., heads),
p = 0.5 (probability of getting heads in a single toss).

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

$$= {}^{n}_{x}C\ P^{x}\ (1 - P)^{n - x}$$

$$= {}^{10}_{3}C\ 0.5^3\ (1 - 0.5)^{10 - 3}$$

$$= 120 \times 0.125 \times 0.00781$$

$$= 0.117$$

$${}^{10}_{3}C = \frac{n!}{x!\ (n-x)!}$$

$$= \frac{10!}{3!\ (10-3)!}$$

$$= 120$$

Example 02

Genetics states that children receive genes from their parents independently. Each child of a particular pair of parents has a probability of 0.3 of having type O blood. This particular pair of parents have 5 children and the number who have type O blood is recorded. What is the probability that exactly 2 of them have type O blood?

There are n = 5 trials, as there are 5 children.
There are two possible outcomes for each child; they either have type O blood or they don't.
The probability that each child has type O blood is p = 0.3.
The trials are independent as children receive genes from their parents independently.

n = 5
x = 2
p = 0.3

$$^{5}_{2}C = \frac{n!}{x!\,(n-x)!}$$

$$= \frac{5!}{2!\,(5-2)!}$$

= 10

$$p(X = 2) = \binom{n}{x} p^{x}(1 - p)^{n-x}$$

$$= {}^{n}_{x}C\,P^{x}\,(1 - P)^{n - x}$$

$$= {}^{5}_{2}C\;\; 0.3^{2}\,(1 - 0.3)^{5 - 2}$$

$$= 10 \times 0.3^{2} \times 0.7^{3}$$

= 0.3087

Example 03

At ABC supermarket, 60% of customers pay by credit card. Find the probability that in a randomly selected sample of ten customers:

a. exactly two pay by credit card,

$$^{10}_{2}C = \frac{n!}{x!\,(n-x)!}$$

$$= \frac{10!}{2!\,(10-2)!}$$

= 45

$$p(X = 2) = {}^{n}_{x}C\ P^{x}\ (1 - P)^{n - x}$$

$$= {}^{10}_{2}C\ \ 0.6^{2}\ (1 - 0.6)^{10 - 2}$$

$$= 45 \times 0.36 \times 0.4^{8}$$

$$= 0.0106$$

Method 2 (from binomial table)

$$p(x \leq 2) = p(x = 0) + p(x = 1) + p(x = 2)$$

$$0.012 = p(x \leq 1) + p(x = 2)$$

$$0.012 - 0.002 = 0.002 + p(x = 2)$$

$$0.01 = p(x = 2)$$

b. more than seven pay by credit card,

$$p(X > 7) = \sum^{10}_{x=8}\ {}^{n}_{x}C\ P^{x}\ (1 - P)^{n - x}$$

$$= \sum^{10}_{x=8}\ {}^{10}_{x}C\ \ 0.6^{x}\ (1 - 0.6)^{10 - x}$$

$$= 0.1672$$

Or

$$p(x > 7) = p(x = 8) + p(x = 9) + p(x = 10)$$

$$= 0.1209 + 0.0403 + 0.00604$$

$$= 0.1673$$

Or

Method 2 (from binomial table)

$$1 - p(x \leq 7) = 1 - 0.833$$

$$= 0.167$$

c. less than three pay by credit card,

$$p(X < 3) = \sum^{2}_{x=0}\ {}^{n}_{x}C\ P^{x}\ (1 - P)^{n - x}$$

$$= \sum^{2}_{x=0}\ {}^{10}_{x}C\ \ 0.6^{x}\ (1 - 0.6)^{10 - x}$$

$$= 0.012$$

Or

$$p(x < 3) = p(x \leq 2)$$

$$= 0.012$$

d. at least two pay by credit card,

$$p(X \geq 2) = \sum_{x=2}^{10} {}^{n}_{x}C \; P^{x} (1 - P)^{n - x}$$

$$= \sum_{x=2}^{10} {}^{10}_{x}C \; 0.6^{x} (1 - 0.6)^{10 - x}$$

$$= 0.9983$$

Or

$$p(x \geq 2) = 1 - p(x \leq 1)$$

$$= 1 - 0.002$$

$$= 0.998$$

e. at most two pay by credit card.

$$p(X \leq 2) = \sum_{x=0}^{2} {}^{n}_{x}C \; P^{x} (1 - P)^{n - x}$$

$$= \sum_{x=0}^{2} {}^{10}_{x}C \; 0.6^{x} (1 - 0.6)^{10 - x}$$

$$= 0.012$$

Or

$$p(x \leq 2) = 0.012$$

Mean & variance of a binomial distribution

If X is a binomial random variable, the mean and variance of X are given by the following formulae

E (x) = n p
V (x) = n p q q = 1 - p

Example: 04

A large bank that issues many loans for General Motors estimates that 70% of the loans are approved within 24 hours of application. If a consumer group takes a random sample of 25 recent General Motors loan applications, what is the probability that

a. at most 20 were approved within 24 hours?

$$p(X \leq 20) = \sum_{x=0}^{20} {}^{n}_{x}C \; P^{x} (1 - P)^{n - x}$$

$$= \sum_{x=0}^{20} {}^{25}_{x}C \; 0.7^{x} (1 - 0.7)^{25 - x}$$

$$= 0.910$$

b. at least 15 were approved within 24 hours?

This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

$$p(X \geq 15) = \sum_{x=15}^{25} {}^{n}_{x}C \; P^{x} (1 - P)^{n - x}$$

$$= \sum_{x=15}^{25} {}^{25}_{x}C \; 0.7^{x} (1 - 0.7)^{25 - x}$$

$$= 0.902$$

Or

$$p(x \geq 15) = 1 - p(x \leq 14)$$

$$= 1 - 0.098$$

$$= 0.902$$

c. at least 12, but no more than 18 are approved within 24 hours?

$$p(12 \leq x < 18) = p(x = 12) + p(x = 13) \ldots\ldots + p(x = 17)$$

$$= \sum_{x=12}^{17} {}^{25}_{x}C \; 0.7^{x} (1 - 0.7)^{25 - x}$$

$$= 0.482$$

Or

0, 1 , 2 , 3 , 4 , 5 , 6 , 7 , 8 , 9 , 10 , 11 , **12 ….. 17** , 18 ……. 25

$$p(x \leq 17) - p(x \leq 11) = p(12 \leq x < 18)$$

$$0.488 - 0.006 = p(12 \leq x < 18)$$

$$0.482 = p(12 \leq x < 18)$$

d. exactly 22 were approved within 24 hours?

$$p(X = 22) = {}^{n}_{x}C \; P^{x} (1 - P)^{n - x}$$

$$= {}^{25}_{x}C \; 0.7^{22} (1 - 0.7)^{25 - 22}$$

$$= 0.0243$$

f. What is the expected number and variance of the loans that will be approved?

$$E(x) = n\,p$$

$$= 25 \times 0.7$$

$$= 17.5$$

$$V(x) = n\,p\,q$$

$$= 25 \times 0.7 \times (1 - 0.7)$$

$$= 25 \times 0.7 \times 0.3$$

$$= 5.25$$

★ **Poisson Distribution** [(Tutorial 04)]

It is used for events that occurs randomly in a specified unit of space, distance or time

Eg:

- ➔ No of phone calls receive in a telephone exchange
- ➔ No of customers entered in a shop in a given time
- ➔ The number of machine breakdowns during a night shift.
- ➔ The number of particles of a particular pollutant in a cubic meter of air emitted from a factory.
- ➔ Number of students arriving during office hours.
- ➔ Number of cars passing through a junction in one minute.
  (This is an example of an interval of time — the time being one minute.)

Assumptions

- ➔ The events must occur independently of each other.
- ➔ The events must occur in a constant rate
  - ◆ i.e number of occurrences in the interval is proportional to the length of the interval
- ➔ The probability of occurring two or more events in a very small interval is insignificant

Poisson formula.

$$P(X = x) = \frac{e^{-\lambda}\ \lambda^{x}}{x!} \quad ; for\ x = 0,1,2 \dots \dots \dots.$$

where λ (λ>0)is the average number of occurrences in a given interval or region and is a constant value of 2.71828… which can be found on your calculator.

Example : 01

On average a person gets 5 phone calls daily. Find the probability that he gets exactly 5 phone calls

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

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

$$= 0.175$$

Example : 02

A paint factory uses “Agent A” in the paint manufacturing process. There is an average of 3 particles of Agent A in a cubic meter of air emitted during the production process. The number of Agent A particles has a Poisson distribution with a mean of 3 particles per cubic meter of air emitted from the factory.

a. What is the probability that there will be 5 particles of Agent A in a cubic meter of air emitted from the factory ?

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

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

$$= 0.1008$$

Method 2 (from Poisson table)

P (x ≤ 5) = p (x = 0) + p (x = 1) + p (x = 2) + p (x = 3) + p (x = 4) + p (x = 5)
P (x ≤ 5) = P (x ≤ 4) + p (x = 5)
0.9161 = 0.8153 + p (x = 5)
0.1008 = p (x = 5)

b. What is the probability that there will be no Agent A particles in a cubic meter of air emission?

$$P(X = 0) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \frac{e^{-3}3^0}{0!}$$

$$= 0.0497$$

c. What is the probability that there will be less than 2 particles of agent A in a cubic meter of air emitted from the factory?

$$P(X < 2) = \sum_{x=0}^{1} \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{1} \frac{e^{-3}3^x}{x!}$$

$$= 0.1991$$

Method 2 (from Poisson table)

P (x < 2) = P(x = 0) + P(x = 1)
= P(x ≤ 1)
= 0.1991

Example : 03

The number of accidents that occur at a busy intersection is Poisson distributed with a mean of 3.4 per week. Find the probability of the following events:

a. Less than three accidents in a week

$$P(X < 3) = \sum_{x=0}^{2} \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{2} \frac{e^{-3.4}3.4^x}{x!}$$

$$= 0.3397$$

Method 2 (from Poisson table)

P (x ≤ 2) = 0.3397

b. Five or more accidents in a week

0, 1, 2, 3, 4, **5, 6, 7, 8 ……..**

$$P(x \geq 5) = p(x = 5) + p(x = 6) \ldots\ldots\ldots$$
$$= 1 - p(x \leq 4)$$
$$= 1 - \sum_{x=0}^{4} \frac{e^{-\lambda}\lambda^{x}}{x!}$$
$$= 1 - \sum_{x=0}^{4} \frac{e^{-3.4}3.4^{x}}{x!}$$
$$= 0.2558$$

Or

Method 2 (from Poisson table)

P (x ≥ 5) = 1 - (p (x ≤ 4))
= 1 - 0.7442
= 0.2560

c. Between two and five

0, 1, 2, **3, 4,** 5 ……..

$$P(2 < x < 5) = p(x = 3) + p(x = 4)$$

$$= \sum_{x=3}^{4} \frac{e^{-\lambda}\lambda^{x}}{x!}$$
$$= \sum_{x=0}^{4} \frac{e^{-3.4}3.4^{x}}{x!}$$
$$= 0.404$$

Method 2 (from Poisson table)

P (2 < x < 5) = P (x ≤ 4) - P (x ≤ 2)
= 0.7742 - 0.3397
= 0.404

d. Greater than or equal to two and less than or equal to five

0, 1, **2, 3, 4, 5**

$$P(2 \leq x \leq 5) = p(x = 2) + p(x = 3) + p(x = 4) + p(x = 5)$$

$$= \sum_{x=2}^{5} \frac{e^{-\lambda} \lambda^{x}}{x!}$$

$$= \sum_{x=2}^{5} \frac{e^{-3.4} 3.4^{x}}{x!}$$

$$= 0.7236$$

Method 2 (from Poisson table)

$$P(2 \leq x \leq 5) = P(x \leq 5) - P(x \leq 1)$$
$$= 0.8705 - 0.1468$$
$$= 0.7237$$

e. Greater than or equal to two and less than five

$$P(2 \leq x < 5) = p(x = 2) + p(x = 3) + p(x = 4)$$

$$= \sum_{x=2}^{4} \frac{e^{-\lambda} \lambda^{x}}{x!}$$

$$= \sum_{x=2}^{4} \frac{e^{-3.4} 3.4^{x}}{x!}$$

$$= 0.5973$$

Method 2 (from Poisson table)

$$P(2 \leq x < 5) = P(x \leq 4) - P(x \leq 1)$$
$$= 0.7442 - 0.1468$$
$$= 0.5974$$

f. No accidents today

For 7 days = 3.4

For 1 day = 3.4 / 7 (λnew)

$$p(x = 0) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

$$= \frac{e^{-3.4/7} \left(\frac{3.4}{7}\right)^{0}}{0!}$$

$$= 0.6153$$

Example : 04

The number of defective items returned each day, over 100 days, to a shop is shown below:

Number of defective items (x)	0	1	2	3	4	> 4
Frequency (f)	50	20	15	10	5	0

Assuming that the number of defective items may be approximated by a Poisson distribution, calculate the value of λ and find the probability that:

a. no defective item is returned on a given day,

$$\lambda = \frac{\sum_i x_i}{\sum f_i}$$

$$= \frac{(50\ x\ 0) + (20\ x\ 1) + \ldots\ldots + (>4\ x\ 0)}{(50 + 20 + \ldots\ldots + 0)}$$

= 100 / 100

= 1

λ = 1 per 100 day

$$p\ (x = 0) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \frac{e^{-1}\,1^{0}}{0!}$$

= 0.3678

b. exactly three defective items are returned on a given day,

$$p\ (x = 3) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \frac{e^{-1}\,1^{3}}{3!}$$

= 0.0613

P (x ≤ 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)

= P(x ≤ 2) + P(x = 3)

0.9810 = 0.9197 + P(x = 3)

0.0613 = P(x = 3)

c. three or more defective items are returned on a given day.

This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

P (x ≥ 3) = P(x = 3) + P(x = 4) + ……..
= 1 - P(x ≤ 2)
= 1 - 0.9197
= 0.0803

d. If the shop owner decided to follow the same procedure over 250 days, what would be the value of λ ? Hence or otherwise, answer the questions above in part (a).

100 ——-> 1

1 per day —-> $\frac{1}{100}$

250 days ——-> $\frac{1}{100}$ X 250

$\lambda = 2.5$

$$p\,(x = 0) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \frac{e^{-2.5}\,2.5^{0}}{0!}$$

= 0.0820

Mean & variance of a poisson distribution

E (x) = λ
V (x) = λ

★ Introduction to statistic (Tutorial 05)

Data

Data are measurements or observations that are collected as a source of information
Eg:

- ➔ The number of people in Sri Lanka
- ➔ the countries where people were born
- ➔ the value of sales of a particular product

Types of Data

1. **Qualitative Data**

   They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated.

   For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative.

2. **Quantitative Data**

   These can be measured and not simply observed. They can be numerically represented and calculations can be performed on them.

For example, data on the number of students playing different sports from your can be classified as quantitative.

Sources of Data

1. **Primary Data**

   These are the data that are collected for the first time by an investigator for a specific purpose.

   Primary data are 'pure' in the sense that no statistical operations have been performed on them and they are original.

   An example of primary data is the Census of Sri Lanka

2. **Secondary Data**

   They are the data that are sourced from someplace that has originally collected it.

   This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form. This information is impure as statistical operations may have been performed on them already.

1. **Primary Data**

   Examples for primary data

   - ➔ Customer surveys
   - ➔ Market research
   - ➔ Scientific experiments
   - ➔ Traffic counts

Primary Data Collection Methods

- ➔ **Interviews**

  It involves two groups of people, where the first group is the interviewer (the researcher(s) asking questions and collecting data) and the interviewee (the subject or respondent that is being asked questions).

  Interviews can be carried out in 2 ways, namely; in-person interviews and telephonic interviews.

  An in-person interview requires an interviewer or a group of interviewers to ask questions from the interviewee in a face-to-face fashion. It can be direct or indirect, structured or structure, focused or unfocused

- ➔ **Surveys & Questionnaires**

  They are a group of questions typed or written down and sent to the sample of study to give responses.
  After giving the required responses, the survey is given back to the researcher to record.

There are 2 main types of surveys used for data collection, namely; online and offline surveys.

➔ **Observation**

The observation method is mostly used in studies related to behavioral science.
There are different approaches to the observation method—structured or unstructured, controlled or uncontrolled, and participant, non-participant, or disguised approach

2. **Secondary data**

Secondary data collection methods

data are available in various resources including

- ➔ Government publications
- ➔ Public records
- ➔ Historical and statistical documents
- ➔ Business documents
- ➔ Technical and trade journals

Sampling

Sampling is the process of selecting units from a population.

Population: it is the set of all observations considered in the research
Sample: sample is a subset of a population

Advantages of sampling

- ➔ Cost is lower
- ➔ Data collection is faster
- ➔ Improve accuracy & quality
- ➔ If the items are destroyed through the test then sampling is the only alternative

Sampling methods

Probability Sampling

In probability sampling technique the units are selected from the population at random using probabilistic methods.

reasons for using probability sampling

- ➔ Making statistical inferences
- ➔ Achieving a representative sample
- ➔ Minimising sampling bias

bias means that the units selected from the population for inclusion in your sample are not representative of that population

Types of probability sampling techniques

➔ Random sampling

In this method, each item in the population has the same probability of being selected as part of the sample as any other item.
Random sampling can be done with or without replacement.
eg: 100 are listed and a group of 20 may be selected from this list at random.

➔ Systematic sampling

In this method, every nth element from the list is selected as the sample.
eg: from a list we would select every 5th, 10th, 15th, 20th, etc

➔ Stratified sampling

A stratum is a subset of the population that share at least one common characteristic. Within each stratum, a simple random sample or systematic sample is selected. Examples of stratums might be males and females, or managers and non-managers.

➔ Cluster Sampling

The population is divided into groups called clusters & a set of clusters are selected randomly to include in the sample.

- One-stage sampling: All of the elements within selected clusters are included in the sample.

- Two-stage sampling: A subset of elements within selected clusters are randomly selected for inclusion in the sample

Non Probability sampling

In non probability sampling the samples are selected based on the subjective judgement of the researcher, rather than random selection

reasons for using non probability sampling

the procedures used to select units for inclusion in a sample are much easier, quicker and cheaper when compared with probability sampling

Types of non probability sampling techniques

➔ Quota sampling

In quota sampling, the aim is to end up with a sample where the strata (groups) being studied (e.g. males vs. females students) are proportional to the population being studied. For instance, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop.

➔ Convenience sampling

A convenience sample is simply one where the units that are selected for inclusion in the sample are the easiest to access. For example if there are 10,000 students, if the sample size is 100 students, we may stand at the main entrance, & gather data from passing by students

➔ Purposive sampling

We sample with a purpose in mind. First the respondents are verified to check whether they meet the criteria for being in the sample. Purposive sampling can be very useful for situations where you need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern.In super market asking questions from certain people

➔ Snowball sampling

In snowball sampling, you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know who also meet the criteria. Snowball sampling is particularly appropriate when the population you are interested in is hidden and/or hard-to-reach. These include populations such as drug addicts, homeless people, individuals with AIDS.

★ **Presentation of Data (Tutorial 06)**

Presentation of data makes it easy to understand about a dataset and help to make correct interpretations

Two methods of presenting data are
- ➔ Tabular form – presenting data in a simple table
- ➔ Pictorial form – data are presented using diagrams, charts or graphs

Tables

➔ **Frequency table**

The frequency of a data value is the number of times the data value occurs

Eg: the marks awarded for an assignment for 20 students are as follows. Present this information in a frequency table

6, 4, 7, 10, 5, 6, 7, 8, 7, 8, 8, 9, 7, 5, 6, 6, 9, 4, 7, 8

Value	Tally	Frequency
4	//	2
5	//	2
6	////	4
7	~~///~~	5
8	////	4
9	//	2
10	/	1

➔ **Cumulative frequency table**

Cumulative frequency is the total of a frequency and all frequencies below it in a frequency distribution

Value	Frequency	Cumulative Frequency
4	2	2
5	2	4
6	4	8
7	5	13
8	4	17
9	2	19
10	1	20

➔ **Group Frequency Table**

When the set of data values are spread out, it is difficult to set up a frequency table for every data value as there will be too many rows in the table. So we group the data into **class intervals**

Eg: The number of calls per day for taxi service was recorded for the month of December. The results were as follows:

28	122	217	130	120	86	80	90	120	140
70	40	145	187	113	90	68	174	194	170
100	75	104	97	75	123	100	82	109	120
81									

Set up a grouped frequency table for this set of data values.

Value	Tally	Frequency
28 - 60	//	2
61 - 93	~~////~~ ~~////~~	10
94 - 126	~~////~~ ~~////~~ /	11
127 - 159	///	3
160 - 192	///	3
192 - 225	//	2

➔ **Relative Frequency and percentage frequency table**

Relative frequency & percentage frequency columns for the above table is as follows

Value	Frequency	Relative frequency	Percentage frequency
28 - 60	2	2 / 31 = 0.06	6
61 - 93	10	10 / 31 = 0.32	32
94 - 126	11	11 / 31 = 0.35	35
127 - 159	3	3 / 31 = 0.09	9
160 - 192	3	3 / 31 = 0.09	9
192 - 225	2	2 / 31 = 0.06	6

$$\text{Relative frequency} = \frac{\text{frequency (f)}}{\text{total frequency}}$$

Definitions

Sales	Number of days
36 - 40	2
41 - 45	7
46 - 50	8
51 - 55	11
56 - 60	2

Consider the example above

➔ **Class interval**

a range into which data may be grouped

eg: 36 – 40

➔ **Class limit**

41 is the lower limit and 45 is the upper limit of the class interval 41-45

➔ **Real limit / class boundary**

40.5 & 45.5 are the real limit of the class 41- 45

Lower real limit = (lower limit of the class + upper limit of the previous class) / 2

Upper real limit = (upper limit of the class + lower limit of the next class) / 2

➔ **Class mark**

class mark of 41 - 45 class is (41+45) / 2 = 43

➔ **Class width**

difference between the real limits. Class width of 41 - 45 class is 45.5 – 40.5 = 5

Graphs

➔ **Histogram**

Steps to draw a histogram

1. Mark frequencies in the ‘y’ axis
2. Mark class boundaries in the ‘x’ axis
3. Determine the height of the rectangle

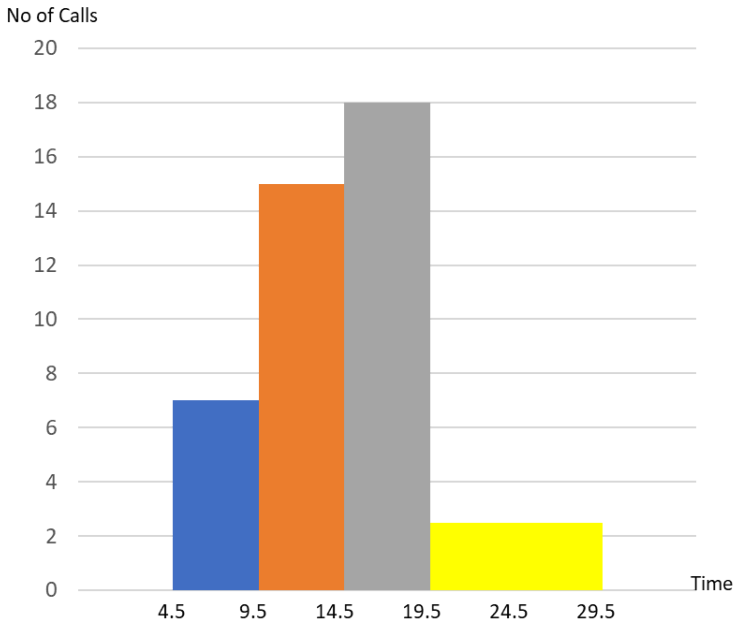
If the class widths are equal then take frequency as the height of the rectangle

If class widths are not equal

$$\text{height} = \frac{\text{frequency}}{\text{class width}} \times k$$ (k is a constant)

Eg: draw a histogram for following data

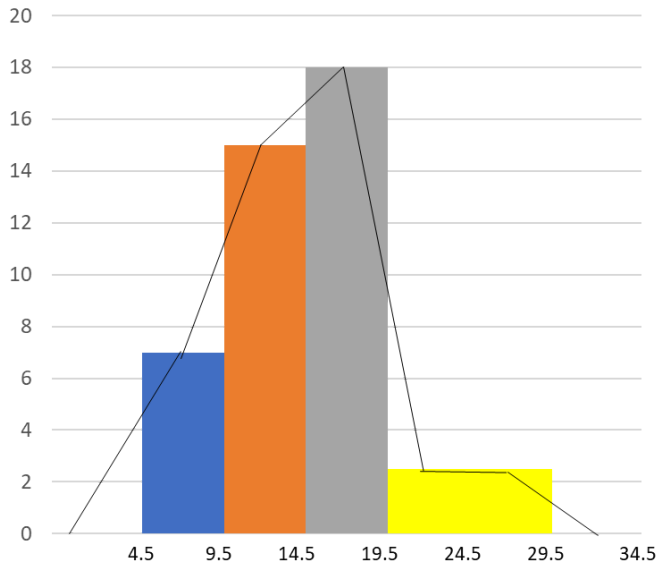
Time (discrete cases)	Continuous cases	Number of calls
5 - 9	4.5 - 9.5	7
10 - 14	9.5 - 14.5	15
15 - 19	14.5 - 19.5	18
20 - 24	19.5 - 24.5	3
25 - 29	24.5 - 29.5	3



➔ **Frequency Polygon**

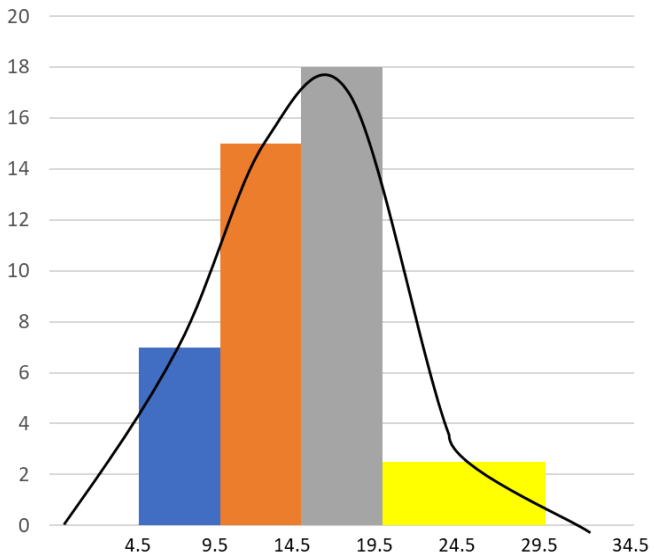
Frequency polygon is a line graph drawn by joining all the mid points of top of the bars of a histogram

Area of the frequency polygon = area of the frequency histogram



➔ **Frequency Curve**

If we fit a smooth curve to a frequency polygon, we get a frequency curve



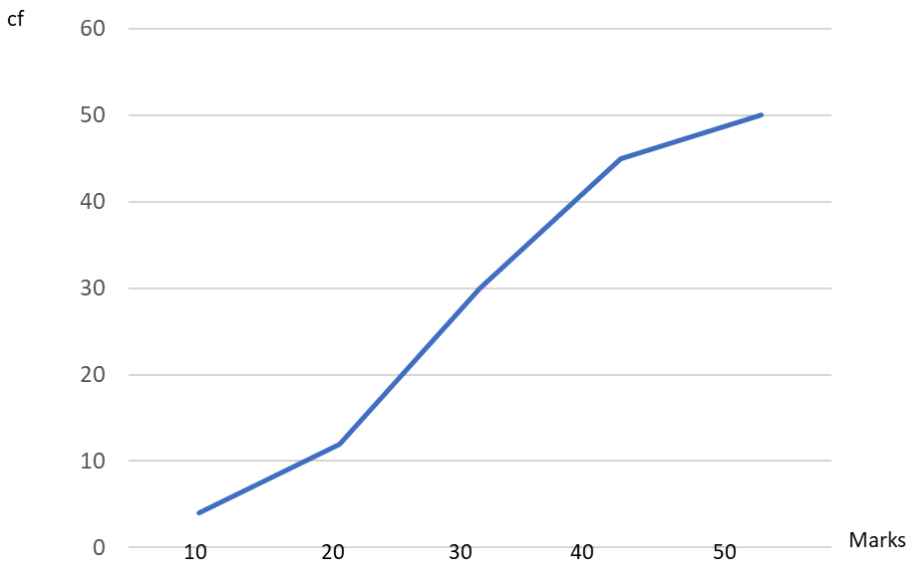
➔ **Cumulative frequency curve / Ogive**

Ogive is the graphical representation of a cumulative frequency distribution

1. **Less than ogive**

- Cumulative frequencies are in the ascending order
- The cumulative frequency of each class is plotted against the upper limit of the class interval

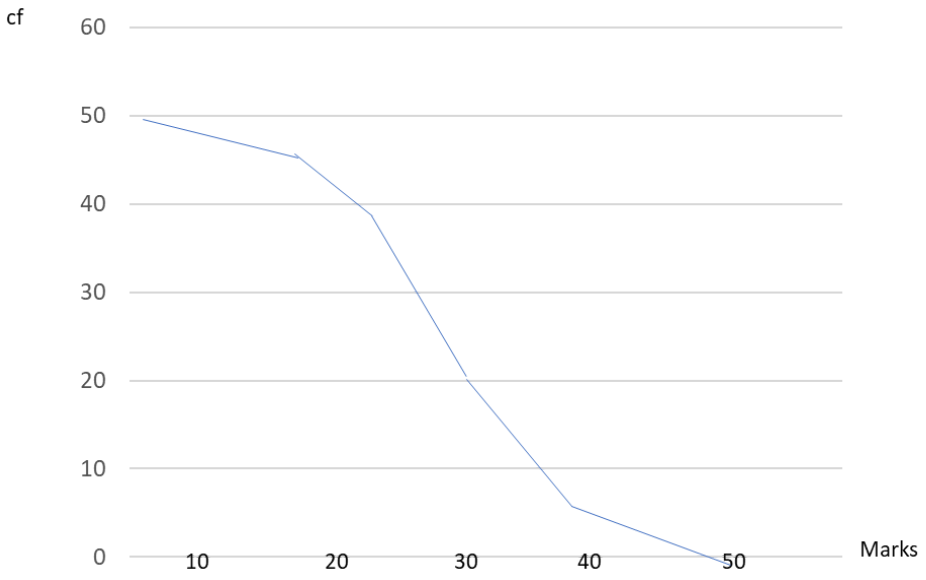
Marks	Number of student	Less than cumulative frequency
0 - 10	4	4
10 - 20	8	12
20 - 30	18	30
30 - 40	15	45
40 - 50	5	50



2. Greater than ogive

- Cumulative frequencies are in the descending order
- The cumulative frequency of each class is plotted against the lower limit of the class interval

Marks	Number of student	Greater than cumulative frequency
0 - 10	4	50
10 - 20	8	46
20 - 30	18	38
30 - 40	15	20
40 - 50	5	5



★ Central Tendency Measures [(Tutorial 07)]

Measures that indicate the central value of a distribution. The 3 most common measures of central tendency are ,

- ➔ Mean
- ➔ Median
- ➔ mode

Mean

Mean is the average value of a data set

Different types of means

- ➔ Arithmetic mean
- ➔ Weighted mean
- ➔ Harmonic mean
- ➔ Geometric mean

Arithmetic Mean

- Mean of ungrouped data

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

The arithmetic mean of the numbers 8, 3, 5, 12, and 10 is

$$\bar{X} = \frac{8 + 3 + 5 + 12 + 10}{5} = \frac{38}{5} = 7.6$$

- mean for ungrouped data (with a frequency table)

Data	Frequency
x_1	f_1
x_2	f_2
.	.
.	.
x_n	f_n

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Example : 01

Number of tracks on album	Frequency
9	1
10	4
11	3
12	3
13	0
14	1

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{(9x1)+(10x4)+(11x3)+(12x3)+(13x0)+(14x1)}{1+4+3+3+0+1}$$

$$= \frac{132}{12}$$

= 11

- mean for grouped data

$$\bar{x} = \frac{\sum^{n}_{i=1} f_i x_i}{\sum^{n}_{i=1} f_i}$$

Where,

x_i = *mid value of the* i th *class interval*

$$= \frac{lower\ limit\ of\ \ i^{th}\ \ class\ interval + upper\ limit\ of\ \ i^{th}\ class\ interval}{2}$$

Example 02

# of patients	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
# of days attending the hospital	2	6	9	7	4	2

Find the average number of patients attending the hospital per day.

# of patients (x)	**Frequency (f)**	**Mid value (x_i)**	**$f_i x_i$**
0 -10	2	(0+10) / 2 = 5	2*5 = 10
10 - 20	6	(10+20) / 2=15	6*15 = 90
20 - 30	9	(20+30) / 2=25	9*25 = 225
30 - 40	7	(30+40) / 2=35	7*35 = 245
40 - 50	4	(40+50) / 2=45	4*45 = 180
50 - 60	2	(50+60) / 2=55	2*55 = 110
	$\sum$ f = 30		$\sum$ f x = 860

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{860}{30}$$

= 28.67

≈ 29 #*of patients per day*

No. of patients cannot be a decimal value

Weighted mean

Find the **weighted mean** of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\overline{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\Sigma wX}{\Sigma w}$$

where $w_1, w_2, \ldots, w_n$ are the weights and $X_1, X_2, \ldots, X_n$ are the values.

Example: 01

A student received an A in English Composition (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology (4 credits), and a D in Physical Education (2 credits). Assuming A = 4 grade points, B = 3 grade points, C = 2 grade points, D = 1 grade point, and F = 0 grade points, find the student's grade point average.

Solution

Course	Credits (w)	Grade (X)
English Composition I	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 point)

$$\overline{X} = \frac{\Sigma wX}{\Sigma w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} = 2.7$$

The grade point average is 2.7.

Median

Median is the middle value of an ordered set of data

- Median of ungrouped data

$$M = \left(\frac{n+1}{2}\right)^{th} term$$

Example : 01

Compute the median for the data:
1.90 3.00 2.53 3.71 2.12 1.76 2.71 1.39 4.00 3.33

Solution
First arrange the given data in numerical order:
1.39 1.76 1.90 2.12 2.53 2.71 3.00 3.33 3.71 4.00

$$M = \left(\frac{n+1}{2}\right)^{th} obs$$

$$M = \left(\frac{10+1}{2}\right)^{th} obs$$

$$M = 5.5^{th} obs$$

$$M = \left(\frac{5^{th} item + 6^{th} item}{2}\right)$$

$$M = \frac{2.53 + 2.71}{2}$$

$$M = \frac{5.24}{2}$$

$$M = 2.62$$

Example: 02

Find the median (with frequency table)

Number of tracks on album	Frequency	Cumulative Frequency (CF)
9	1	1
10	4	5
11	3	8
12	3	11
13	0	11
14	1	12

$$M = \left(\frac{n+1}{2}\right)^{th} term$$

$$M = \left(\frac{12+1}{2}\right)^{th} term$$

$$M = 6.5^{th}$$

$$Median = 11$$

- Median of grouped frequency distributions

$$M_d = L + \frac{\left(\frac{n}{2} - c\right)}{f} x\ h$$

where

L - lower real limit of the median class,
C - cumulative frequency preceding the median class
f - frequency of the median class
h – width of the median class
n – number of data

Example : 01

Height		Frequency	Cumulative Frequency (CF)
60 - 62	59.5 - 62.5	5	5
63 - 65	62.5 - 65.5	18	**23**
66 - 68	**65.5 - 68.5**	**42**	65
69 - 71	68.5 - 71.5	27	92
72 - 73	71.5 - 73.5	8	100

$$M = \left(\frac{n+1}{2}\right)^{th} obs$$
$$M = \left(\frac{100+1}{2}\right)^{th} obs$$
$$M = 50.5^{th}$$
$$M = 66 - 68$$

$$M_d = L + \frac{\left(\frac{n}{2} - c\right)}{f} x\, h$$
$$M_d = 65.5 + \frac{\left(\frac{100}{2} - 23\right)}{42} x\, 3$$
$$M_d = 67.42$$

Mode [(Tutorial 08)]

Mode is the most repeated term in a group of data.
The mode of a set of sample data is the most frequently occurring value.

➔ Unimodal

A data set that has only one value that occurs with the greatest frequency is said to be unimodal.

Mode of 3,7,9,4,5,6,3,7,1,7 is 7

➔ Bimodal

If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be bimodal.

There are 2 modes for the data set 2,4,2,4,5,4,7,8,9,8,8. They are 4 and 8. these kind of distributions are called bimodal distributions

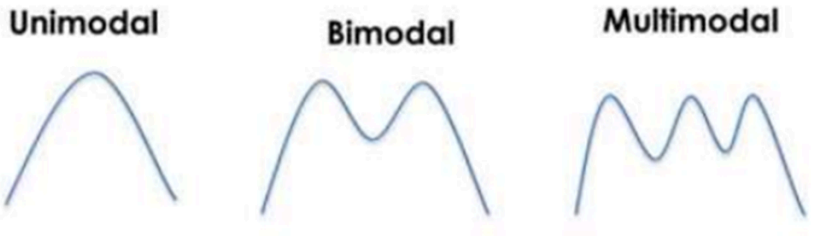
➔ Multimodal

If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be multimodal.

➔ No mode

When no data value occurs more than once, the data set is said to have no mode.

There is no mode for data set 3,5,8,7,9,4,1



- Mode of ungrouped frequency distribution

Number of tracks on album	Frequency
9	1
10	4
11	3
12	3
13	0
14	1

Mode is = 10

- Mode of grouped frequency distribution

$$M = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c$$

L - lower boundary of the modal class

Δ_1 - difference between the frequency of the modal class and the class preceding it

Δ_2 - difference between the frequency of the modal class and the class after it

C - class interval of the modal class

This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

Example: 01

Time		Frequency
1 - 10	0.5 - 10.5	**8**
11 - 20	**10.5 - 20.5**	**14**
21 - 30	20.5 - 30.5	**12**
31 - 40	30.5 - 40.5	9
41 - 50	40.5 - 50.5	7

$$M = L + (\frac{\Delta_1}{\Delta_1 + \Delta_2})c$$

$$M = 10.5 + (\frac{14-8}{(14-8) + (14-12)})10$$

$$M = 10.5 + (\frac{6}{6+2})10$$

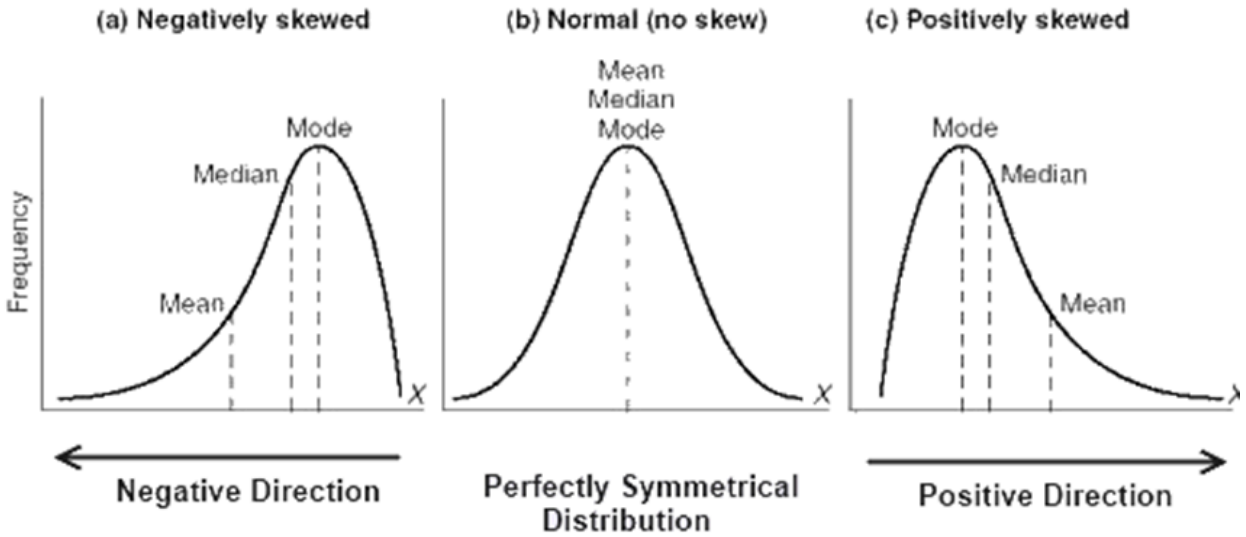
$$M = 10.5 + (\frac{6}{8})10$$

$$M = 10.5 + (0.75)\ x\ 10$$

$$M = 10.5 + 7.5$$

$$M = 18$$

Relationship between mean, median, mode



This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj

★ Measures of dispersion (Tutorial 09)

The degree to which numerical data tends to spread about an average value is called variation or dispersion of data.

Measures of dispersion (or spread) – This measures how far the data is spread around the mean.

The measures of dispersion are,

- ➔ Range
- ➔ Mean deviation
- ➔ Quartile deviation
- ➔ Variance and standard deviation
- ➔ Coefficient of variation

Range

The difference between the largest and the smallest numbers in the data set
If the range is low, then the variability / dispersion is low,

$$R = x_{max} - x_{min}$$

Eg: for the data set 2, 5, 6, 7, 1, 9, 2, 3 range is 8

Variance and standard deviation

- For ungrouped data **(If ask sample variance divide formula to the (n - 1))**

Variance

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Sample Variance

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

standard deviation

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Example : 01

find the variance of 75, 80, 82, 87, 96

$$\text{(mean) } \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$\bar{x} = \frac{(75+80+82+87+96)}{5}$$

$$= 84$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
75	75 - 84 = - 9	81
80	80 - 84 = - 4	16
82	82 - 84 = - 2	4
87	87 - 84 = 3	9
96	96 - 84 = 12	144
		254

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$\sigma^2 = \frac{254}{5}$$

$$\sigma^2 = 50.8 \text{ (variance)}$$

$$\sigma = 7.13 \text{ (standard deviation)}$$

- For ungrouped frequency distribution **(If ask sample variance divide formula to the (f - 1)**

Variance

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}$$

standard deviation

$$\sigma = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}}$$

Example : 02

Find the variance

x	1	2	3	4	5	6	7
f	5	12	8	3	0	0	1

x	f	fx	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f(x_i - \bar{x})^2$
1	5	5	1 - 2.48 = -1.48	2.2	11
2	12	24	- 0.48	0.2	2.4
3	8	24	0.52	0.3	2.4
4	3	12	1.52	2.3	7
5	0	0	2.52	6.4	0
6	0	0	3.52	12.4	0
7	1	7	4.52	20.4	20.4
	$\sum f$ = 29	$\sum fx$ = 72			43.2

$$\text{(mean) } \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$\bar{x} = \frac{72}{29}$$

$$= 2.48$$

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}$$

$$\sigma^2 = \frac{43.2}{29}$$

$$\sigma^2 = 1.49 \text{ (variance)}$$

- For grouped frequency distribution **(If ask sample variance divide formula to the (f - 1))**

Variance

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}$$

standard deviation

$$\sigma = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}}$$

Example : 03

The following table gives the age of people at a birthday party. Calculate standard deviation.

age	15 - 19	20 - 24	25 - 29	30 - 34	35 -39	40 - 44	45 - 49
# of people	2	15	35	28	11	6	3

Age	f	x_i	fx	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f(x_i - \bar{x})^2$
15 -19	2	17	34	- 13.05	170.30	340.6
20 - 24	15	22	330	- 8.05	64.80	972
25 -29	35	27	945	- 3.05	9.30	325.5
30 - 34	28	32	896	1.95	3.80	106.4
35 - 39	11	37	407	6.95	48.30	531.3
40 - 44	6	42	252	11.95	142.80	856.8
45 - 49	3	47	141	16.95	287.30	861.9
	$\sum f$ = 100		$\sum fx$ = 3005			3994.5

$$\text{(mean) } \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$\bar{x} = \frac{3005}{100}$$

$$= 30.05$$

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}$$

$$\sigma^2 = \frac{3994.5}{100}$$

$$\sigma^2 = 39.945 \text{ (variance)}$$

$$\sigma = 6.32 \text{ (standard deviation)}$$

Coefficient of variation (cv) [(Tutorial 10)]

It is also known as the relative standard deviation

The coefficient of variation (CV) is a relative measure of variability that indicates the size of a standard deviation in relation to its mean

The coefficient of variation (CV) is the ratio of the standard deviation to the mean.

$$cv = \frac{standard\ deviation\ (\sigma)}{mean\ (\bar{x})}\ x\ 100\ \%$$

Low CV: Indicates **high consistency** and **low variability.**
High CV: Indicates **low consistency** and **high variability.**

Higher values indicate that the standard deviation is relatively large compared to the mean. It helps to compare two data sets on the basis of the degree of variation.

Example: 01

Two plants C and D of a factory show the following results about the number of workers and the wages paid to them. Using coefficients of variation formulas, find in which plant, C or D is there greater variability in individual wages.

No. of workers	5000	6000
Average monthly wages	$ 2500	$ 2500
Standard deviation	9	10

$$cv_C = \frac{standard\ deviation\ (\sigma)}{mean\ (\bar{x})}\ x\ 100\ \%$$
$$cv_C = \frac{9}{2500}\ x\ 100\ \%$$
$$cv_C = 0.36\ \%$$

$$cv_D = \frac{standard\ deviation\ (\sigma)}{mean\ (\bar{x})}\ x\ 100\ \%$$
$$cv_D = \frac{10}{2500}\ x\ 100\ \%$$
$$cv_D = 0.40\ \%$$

Hence plant D has greater variability in individual wages.
because its Coefficient of Variation is higher.

Example: 02

The following table gives the values of mean and variance of two variables; heights and weights of G.C.E.(O/L) students of a certain school:

	Height	Weight
Mean	155 cm	46.5 kg
Variance	72.25 cm^2	28.09 kg^2

Which variable has **more variability** than the other? Justify your answer.

$$cv_H = \frac{standard\ deviation\ (\sigma)}{mean\ (\bar{x})}\ x\ 100\ \%$$
$$cv_H = \frac{\sqrt{72.25\ cm2}}{155\ cm}\ x\ 100\ \%$$
$$cv_H = \frac{8.5\ cm}{155\ cm}\ x\ 100\ \%$$
$$cv_H = 5.483\ \%$$

$$cv_w = \frac{standard\ deviation\ (\sigma)}{mean\ (\bar{x})}\ x\ 100\ \%$$
$$cv_w = \frac{\sqrt{28.09\ \ kg2}}{46.5\ kg}\ x\ 100\ \%$$
$$cv_w = \frac{5.3\ kg}{46.5\ kg}\ x\ 100\ \%$$
$$cv_w = 11.397\ \%$$

The weight has more variability than the height.
The coefficient of variation (11.40%) is higher than that of height (5.48%).

Example: 03

The total marks scored by two students Sathya and Vidhya in 5 subjects are 460 and 480 with standard deviation 4.6 and 2.4 respectively. Who is **more consistent in performance ?**

Low CV: Indicates **high consistency** and **low variability.**
High CV: Indicates **low consistency** and **high variability.**

$$cv_{Sathya} = \frac{standard\ deviation\ (\sigma)}{mean\ (\bar{x})}\ x\ 100\ \%$$
$$cv_{Sathya} = \frac{4.6}{92}\ x\ 100\ \%$$
$$cv_{Sathya} = 5\ \%$$

$$cv_{Vidhya} = \frac{standard\ deviation\ (\sigma)}{mean\ (\bar{x})}\ x\ 100\ \%$$
$$cv_{Vidhya} = \frac{2.4}{96}\ x\ 100\ \%$$
$$cv_{Vidhya} = 2.5\ \%$$

Vidhya's performance is more consistent (less variability) because her CV is lower.

Quartiles

- For ungrouped data

$$Q_i = \frac{i\ (n+1)}{4}$$

n = total frequency
i = i th quartile

Example: 01
Consider the data set of GPAs given in the previous example. Find the quartiles of the data set.

We first list the data in numeric order.

1.39 1.76 1.90 2.12 2.53 2.71 3.00 3.33 3.71 4.00

$$Q_i = \frac{i\,(n+1)}{4}$$
$$Q_1 = \frac{1\,(10+1)}{4}$$
$$Q_1 = \frac{11}{4}$$
$$Q_1 = 2.75^{\,th}$$
$$Q_1 = 2^{\,th} + 0.75\,(3^{\,rd} - 2^{\,nd})$$
$$Q_1 = 1.76 + 0.75\,(1.90 - 1.76)$$
$$Q_1 = 1.76 + 0.105$$
$$Q_1 = 1.865$$

$$Q_i = \frac{i\,(n+1)}{4}$$
$$Q_3 = \frac{3\,(10+1)}{4}$$
$$Q_3 = \frac{33}{4}$$
$$Q_3 = 8.25^{\,th}$$
$$Q_3 = 8^{\,th} + 0.25\,(9^{\,th} - 8^{\,th})$$
$$Q_3 = 3.33 + 0.25\,(3.71 - 3.33)$$
$$Q_3 = 3.33 + 0.095$$
$$Q_3 = 3.425$$

$$IQR = Q_3 - Q_1$$
$$IQR = 3.425 - 1.865$$
$$IQR = 1.56$$

- For grouped data

$$Q_i = \frac{i\,N}{4}$$

N = total frequency

i = i th quartile

Example: 01

Class	Frequency	cf
1 - 5	6	6
5 - 10	19	25
10 - 15	13	38
15 - 20	20	58
20 - 25	12	70
25 - 30	11	81
30 - 35	6	87
35 - 40	5	92

$$Q_i = \frac{i\,N}{4}$$
$$Q_1 = \frac{1\,x\,92}{4}$$
$$Q_1 = 23^{th}$$
$$class = 5 - 10$$

$$Q_i = \frac{i\,N}{4}$$
$$Q_2 = \frac{2\,x\,92}{4}$$
$$Q_2 = 46^{th}$$
$$class = 15 - 20$$

$$Q_3 = \frac{i\,N}{4}$$
$$Q_3 = \frac{3\,x\,92}{4}$$
$$Q_3 = 69^{th}$$
$$class = 20 - 25$$

Inter quartile Range (IQR)

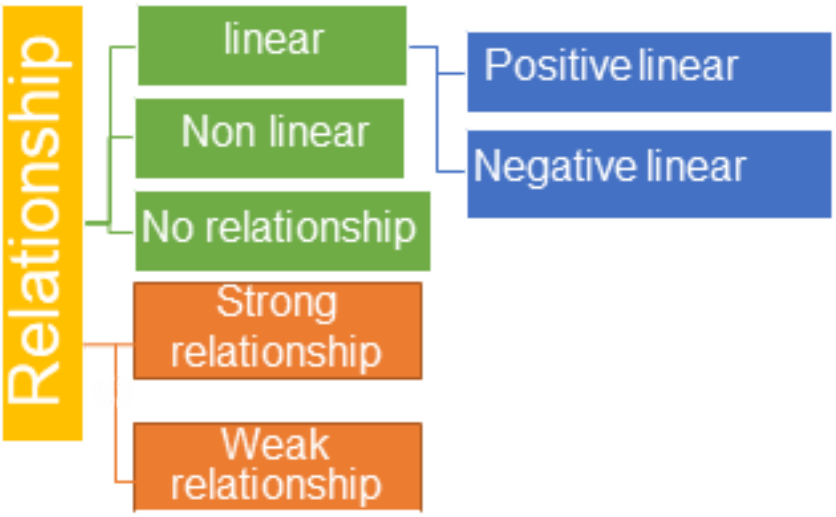
$$IQR = Q_3 - Q_1$$

★ Correlation Analysis (Tutorial 11)

Correlation

Correlation is the relationship between two variables.
Correlation analysis measures the strength of the relationship between two variables.

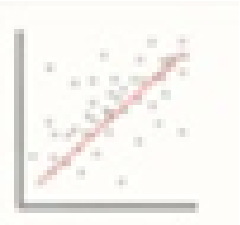
Types of relationships



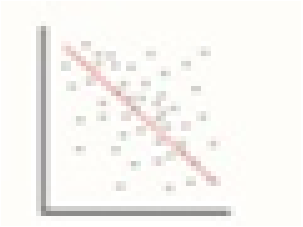
➔ Linear relationship

A relationship between X and Y is linear if a straight line provides an adequate representation of the average relationship between the two variables

1. Positive relationship – a relationship between X and Y is positive if increasing in X tend to be associated with increases in Y and decreasing in X tend to be associated with decreases in Y



2. Negative relationship – the relationship between X and Y is negative if increasing in X tend to be associated with decreases in Y and decreasing in X tend to be associated with increase in Y



➔ **Non linear relationships**

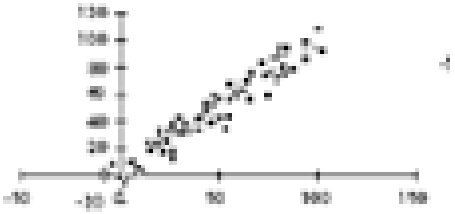
A relationship where changes in the Y value do not change in direct proportion to changes in the X value

➔ **No relationship**

` A relationship in which change in X do not seem to have any effect on the variable of Y

➔ **Strong relationship**

The relationship between X and Y is relatively strong if the points in the scatter diagram lie close to the line of average relationship



➔ **Weak relationship**

The relationship between X and Y is minimal

Spearmen's Rank Correlation Coefficient

It is a measure of the strength of a linear relationship between two variables

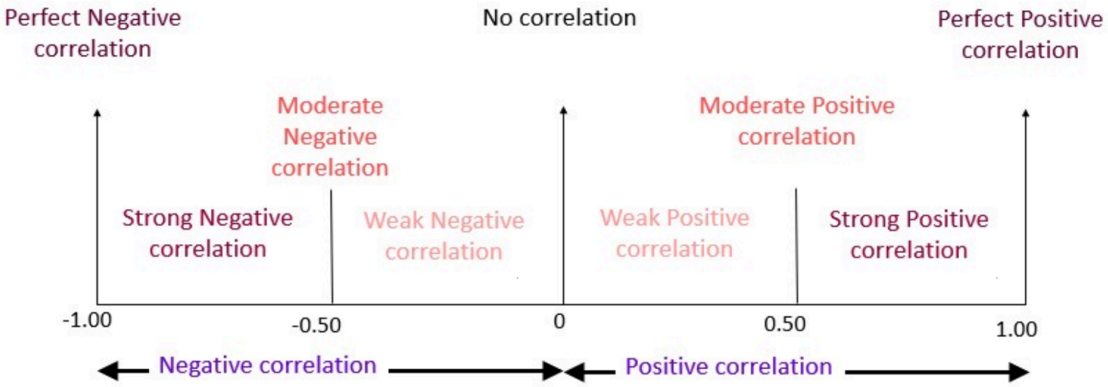
$$P = 1 - \frac{6 \sum d_i^2}{n\,(n^2 - 1)}$$

P = Spearmen's rank correlation coefficient

d_i = different between the two ranks of each observation

n = number of observation

When p = 1 there is a perfect positive correlation
When p = 0 there is no linear correlation between the variables
When p = - 1 there is a perfect negative correlation



Example: 01

ranks of 8 students according to the marks obtained for mathematics & statistics are shown below. Find the rank correlation coefficient & interpret the results

Student	A	B	C	D	E	F	G	H
Maths	3	8	1	6	7	2	5	4
Stat	2	7	3	5	8	1	6	4

Student	Maths	Stat	d	d^2
A	3	2	1	1
B	8	7	1	1
C	1	3	- 2	4
D	6	5	1	1
E	7	8	-1	1
F	2	1	1	1
G	5	6	-1	1
H	4	4	0	0
				10

$$P = 1 - \frac{6 \sum d_i^2}{n\,(n^2 - 1)}$$

$$P = 1 - \frac{6\ x\ 10}{8\,(8^2 - 1)}$$

$$P = 0.881$$

There is a **strong positive correlation** between maths marks and stats marks. It indicates that a student who is good in maths is also good in stats

Example: 02

The marks of 8 students in English and Tamil language tests were as follows. Calculate rank correlation coefficient

Student	A	B	C	D	E	F	G	H
English	52	25	86	33	56	57	54	46
Tamil	40	48	65	57	41	39	63	34

Steps to Calculate the Rank Correlation Coefficient:

- Start by ranking the English and Tamil marks separately.
- Rank from **highest to lowest.** The highest mark gets rank 1, the second-highest gets rank 2, and so on.
- Whether you Rank in **ascending or descending order**, the result will be the same for Spearman's rank correlation coefficient

Student	English marks	Rank in English	Tamil marks	Rank in Tamil	d_i	d_i^2
A	52	5	40	6	-1	1
B	25	8	48	4	4	16
C	86	1	65	1	0	0
D	33	7	57	3	4	16
E	56	3	41	5	- 2	4
F	57	2	39	7	- 5	25
G	54	4	63	2	2	4
H	46	6	34	8	- 2	4
						70

$$P = 1 - \frac{6 \sum d_i^2}{n\,(n^2 - 1)}$$

$$P = 1 - \frac{6\ x\ 70}{8\,(8^2 - 1)}$$

$$P = 0.166$$

weak positive correlation between English and Tamil marks.

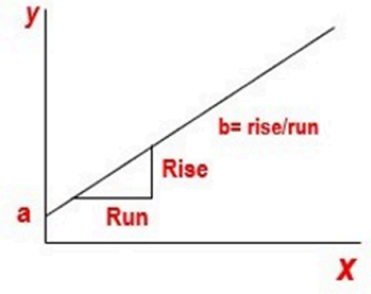
★ **Regression Analysis** [(Tutorial 12)]

Linear regression analysis

Linear regression attempts to model the relationship between two variables by fitting a line to observe data

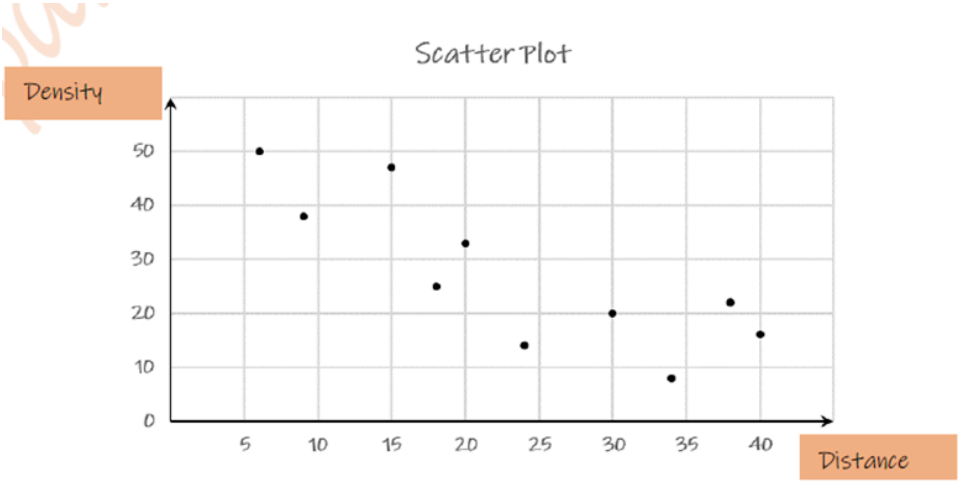
$$Y = a + bx$$

Y = dependent variable
x = independent variable
a = Y-intercept
b = slope of the line



Scatter diagram

It is a mathematical diagram which uses cartesian coordinates to displaying values for two variables for a set of data



- ➔ Independent variable – independent variable is one which is not affected by the changes in other variables. It is plotted along the x axis

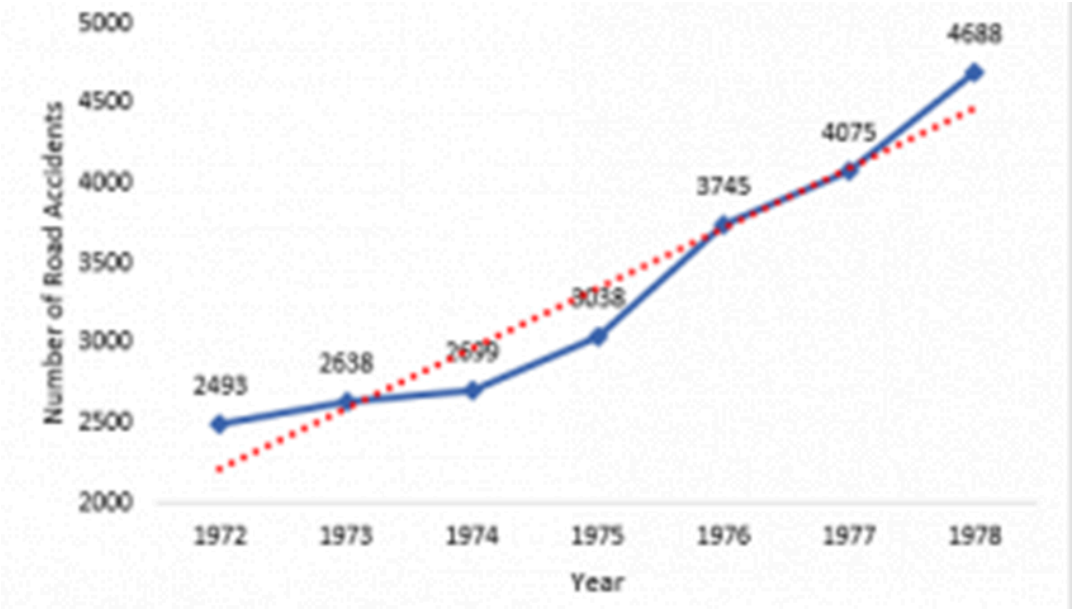
- ➔ Dependent variable – dependent variable is one whose values are determined by the values of the independent variable. It is plotted along y axis

Curve fitting

Methods of curve fitting

- ➔ Free hand method
- ➔ Method of semi average
- ➔ Moving average method
- ➔ Least square method

- ➔ **Free hand method**
  Draw a free hand smooth curve (or a straight line) through the points



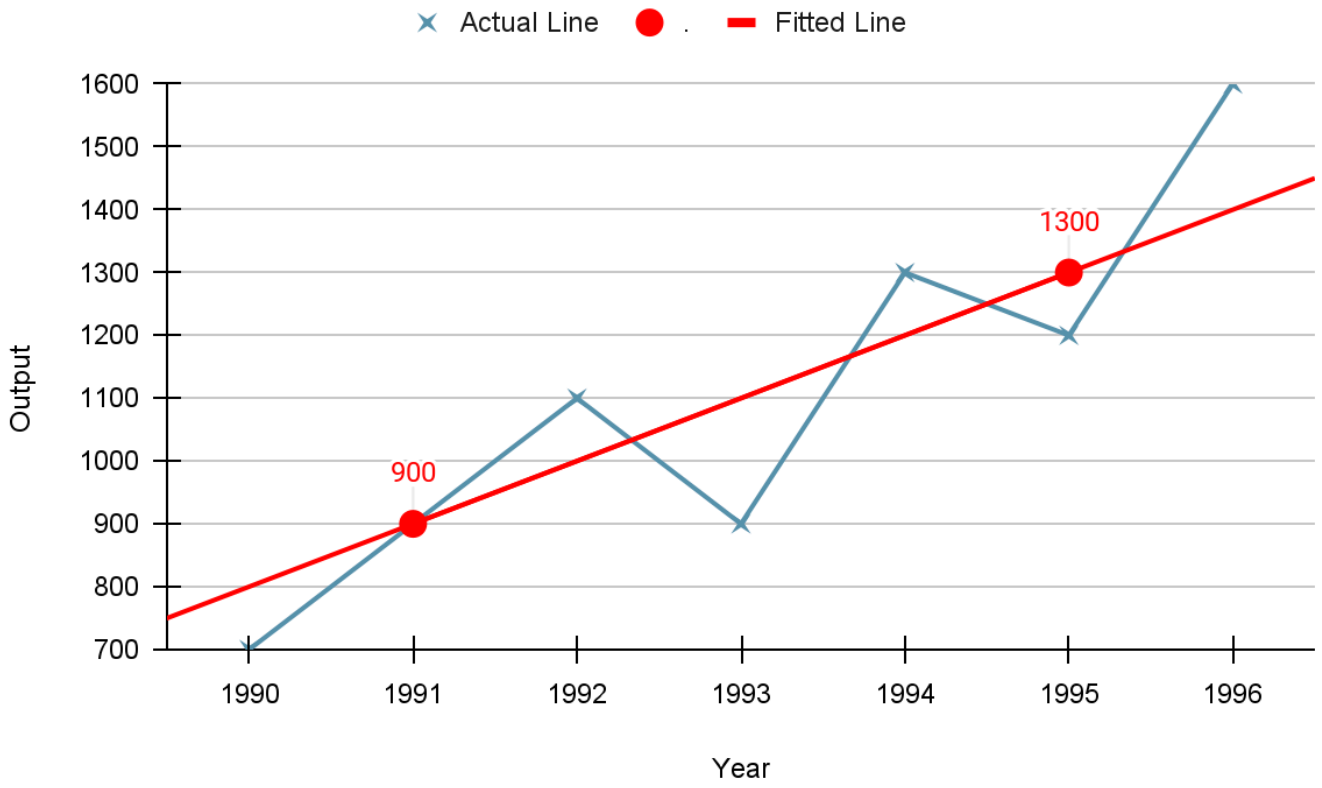
➔ **Method of semi average**

- The data is divided into two equal parts. In case of odd number of data, two equal parts can be made simply by omitting the middle value.
- The average of each part is calculated, thus we get two points.
- Each point is plotted at the mid-point of each half.
- Join the two points by a straight line.
- The straight line can be extended on either side.

Example: 01

Fit a trend line by the method of semi-averages for the given data

Year	1990	1991	1992	1993	1994	1995	1996
Output	700	900	1100	900	1300	1000	1600
Average	$\frac{700+900+1100}{3} = 900$			~~900~~ X	$\frac{1300+1000+1600}{3} = 1300$		

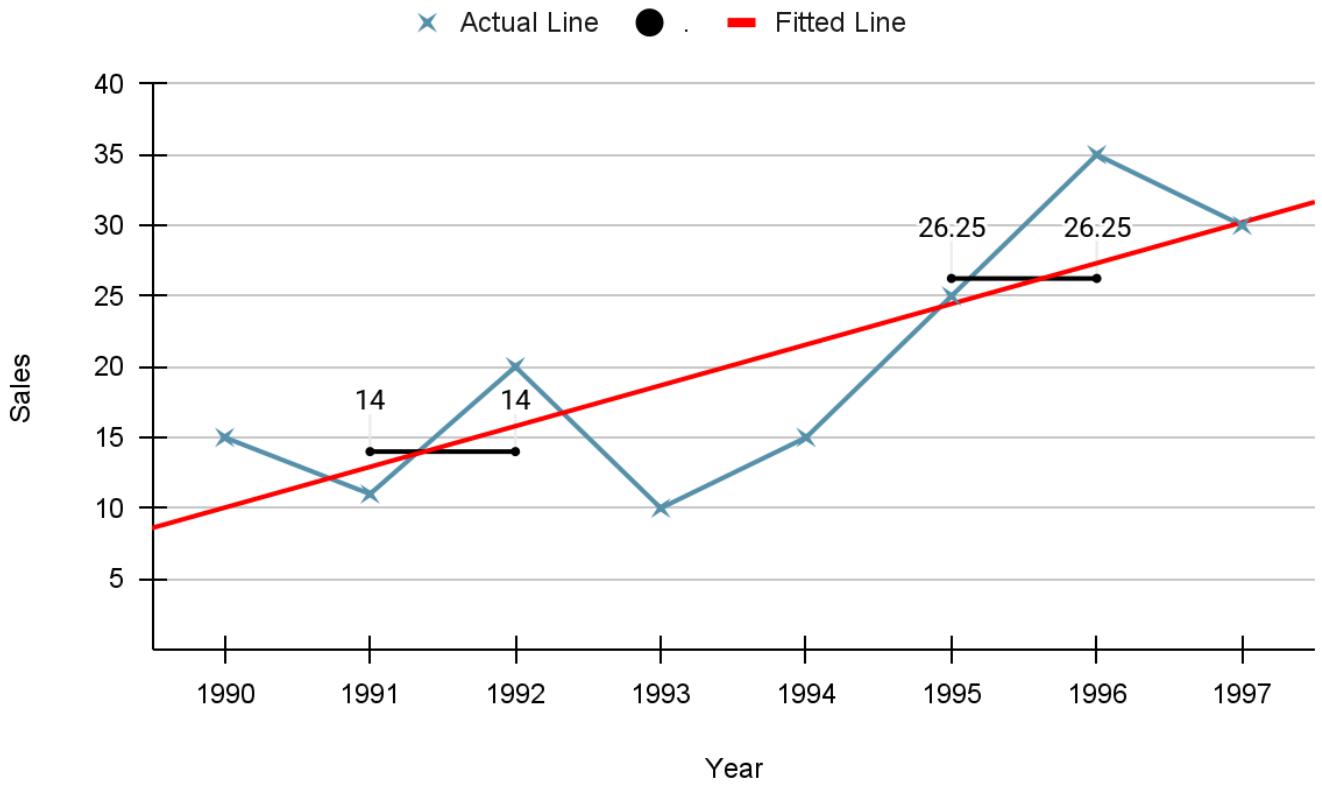


Example: 02

Fit a trend line by the method of semi-averages for the given data

Year	1990	1991	1992	1993	1994	1995	1996	1997
Sales	15	11	20	10	15	25	35	30
Average	$\frac{15+11+20+10}{4} = 14$				$\frac{15+25+35+30}{4} = 26.25$			

This short note making resources are,google search engine,chat gpt and lecture slides
WhatsApp group (click on this/new syllabus notes only) Credit - Kanishka Viraj



➔ **Moving average method** [(Tutorial 13)]

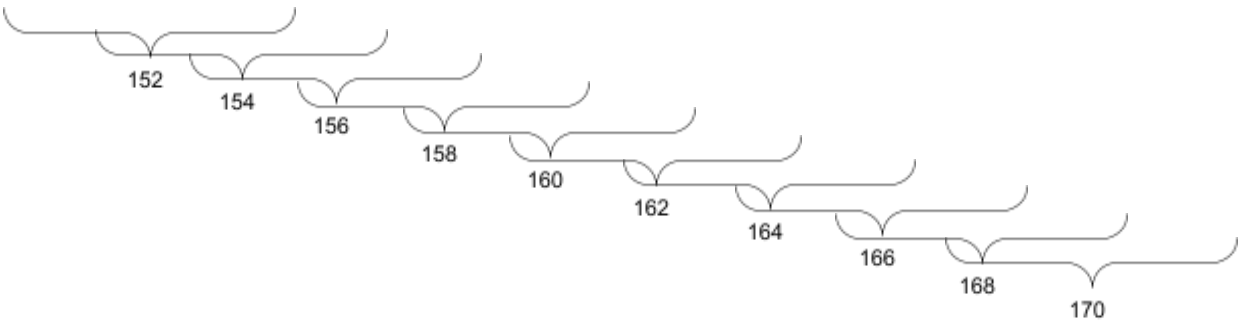
A moving average is a series of averages, calculated from historic data.

Moving averages can be calculated for any number of time periods,
for example a three-month moving average, a seven-day moving average, or a four-quarter moving average.

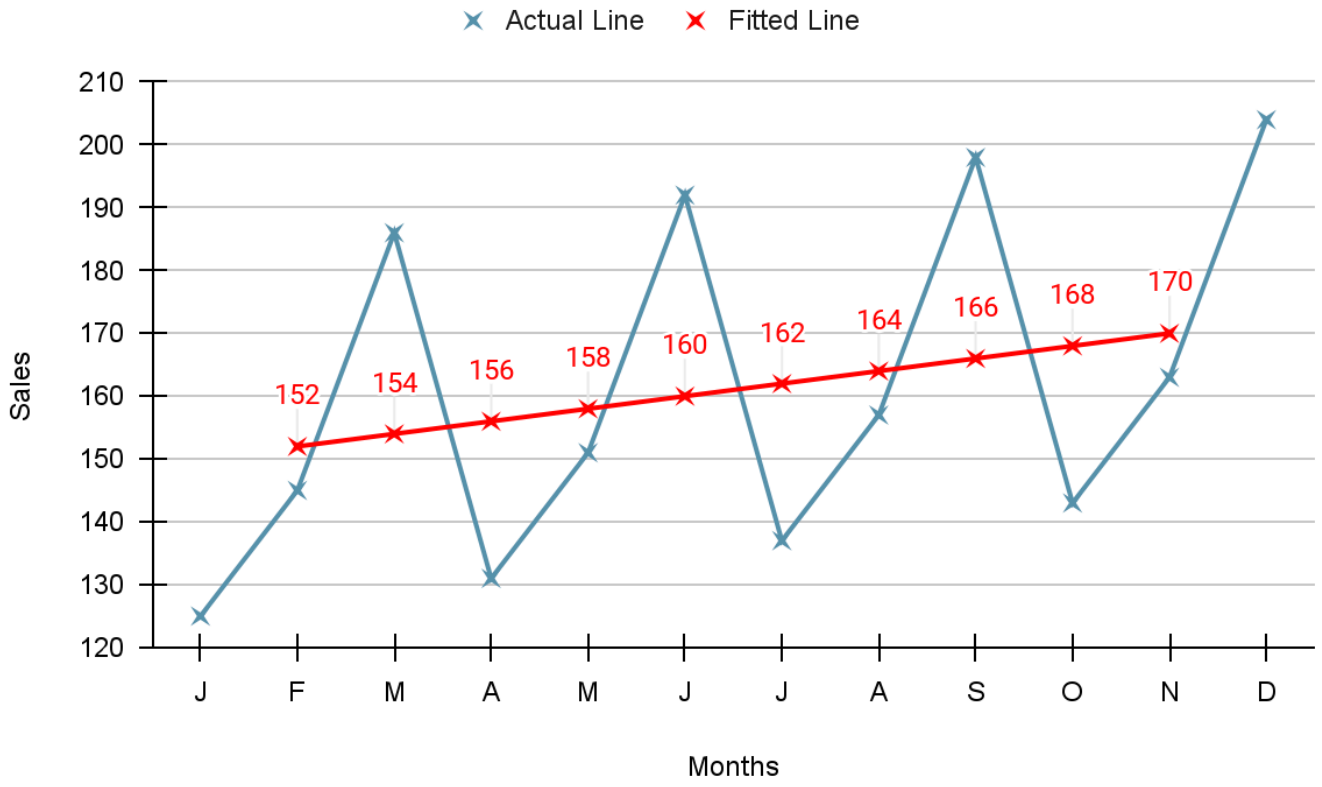
Example: 01

from the data below calculate 3 month moving averages then plot the data and draw the trend line

Months	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
Sales	125	145	186	131	151	192	137	157	198	143	163	204



$\frac{125+145+186}{3} = 152$



➔ **Least square method**

If the least squares regression line y on x is y = a + bx, the values of a and b are found by solving the simultaneous equations.

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

The regression line can be used for estimation, prediction or forecasting.

Example: 01

Using the least square method, find the regression line for the following information.

X	Y
2	4
3	6
7	10
8	12
10	13

X	Y	XY	X^2
2	4	8	4
3	6	18	9
7	10	70	49
8	12	96	64
10	13	130	100
∑x = 30	∑y = 45	∑xy = 322	$\sum x^2 = 226$

$$\Sigma y = na + b\Sigma x$$

45 = 5a + 30b

5a + 30b = 45 —–> (01)

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

322 = 30a + 226b

30a + 226b = 322 —–> (02)

$$a = \frac{51}{23} \quad and \quad b = \frac{26}{23}$$

Therefore, the regression line is y = a + bx

$$y = \frac{51}{23} + \frac{26}{23}x$$

★ Application of the statistics

Introduction

Statistics is indispensable for decision-making in various sectors.

The goal of statistics is to gain understanding from the data. It has a wide range of applications in sectors such as,

- ➔ Health
- ➔ Business and finance
- ➔ Social science
- ➔ IT Industry
- ➔ Banking & insurance
- ➔ Environmental science

Statistical applications in health sector

- Epidemiology

  Epidemiology is the study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine.

- Clinical research

  Clinical research is a branch of healthcare science that determines the safety and effectiveness (efficacy) of medications, devices, diagnostic products and treatment regimens intended for human use.

- Quantitative psychology

  Quantitative psychology is the science of statistically explaining and changing mental processes and behaviors in humans.

Business & finance

- Business analytics

  Business analytics is a rapidly developing business process that applies statistical methods to data sets (often very large) to develop new insights and understanding of business performance & opportunities

- Econometrics

  Econometrics is a branch of economics that applies statistical methods to the empirical study of economic theories and relationships.

- Actuarial science

  Actuarial science is the discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries

Environmental science

- Environmental statistics

  Environmental statistics is the application of statistical methods to environmental science. Weather, climate, air and water quality are included, as are studies of plant and animal populations.

- Population ecology

  Population ecology is a subfield of ecology that deals with the dynamics of species populations and how these populations interact with the environment.

Social science

- Social statistics

  Social statistics is the use of statistical measurement systems to study human behavior in a social environment.

- Psychometrics

  Psychometrics is the theory and technique of educational and psychological measurement of knowledge, abilities, attitudes, and personality traits.

- Demography

  Demography is the statistical study of all populations.