

- We emphasize on data type because tools are used accordingly.

SPSS (software)  
(statistical Package for social sciences)

- We have to decide which tools to use for which type of data.
- Presentation of data is very important.

### ≡ How to present your Data? ≡

#### Presentation.

##### Tabular

- frequency distribution (method)  
many more but fdis one  
of them

##### Graphical / Pictorial

- Bar chart (method)
  - simple
  - multiple
  - component
- Pie chart.
- Histogram.

- We will only look for most used methods.

## - Frequency Distribution.

• consider data is quantitative of age.

• First we will make classes.

↳ Rules:

→ expert opinion.

• Class must be from 6,7 -  $\{ \text{no. of bins} \}$

1) classes  $> 7$  make size of classes accordingly.

• class size can be fix or not.

• Take that class size is  $\frac{\text{size}}{\text{no. of classes}}$  i.e. 6-7 here.

$$\left( \text{size} = \frac{\text{max} - \text{min}}{6} \right) \text{ common algo.}$$

Classes(Age)	C.F	Frequencies	%.
(openend) 0 - 10 (closeend)	5	5	12.20
10 - 20	12	7	
20 - 30	27	15	36.5
30 - 40	35	8	
40 - 50	38	3	
50 - 60	40	2	
60 above .	41	1	must .

→ total data = sum

• Extreme observations can also impact data.

0 - 9

10 - 19

20 - 29

X wrong approach  
disturbs continuity



15 se kam waley:

C.F. → Cumulative frequency.

Sum = Previous frequency.

↳ It can be reverse ~~order~~.

↳ 15 se zaya waley.

→ F.d. for Qualitative data:

Gender	
Male	15
<u>Female</u>	<u>5</u>
	20

→ We can't make classes here like  
1-2 ? is useless. We have to  
represent it in categories.

If categories exceed size, so we ~~can't~~  
divide it.

Cities of Pak ✓

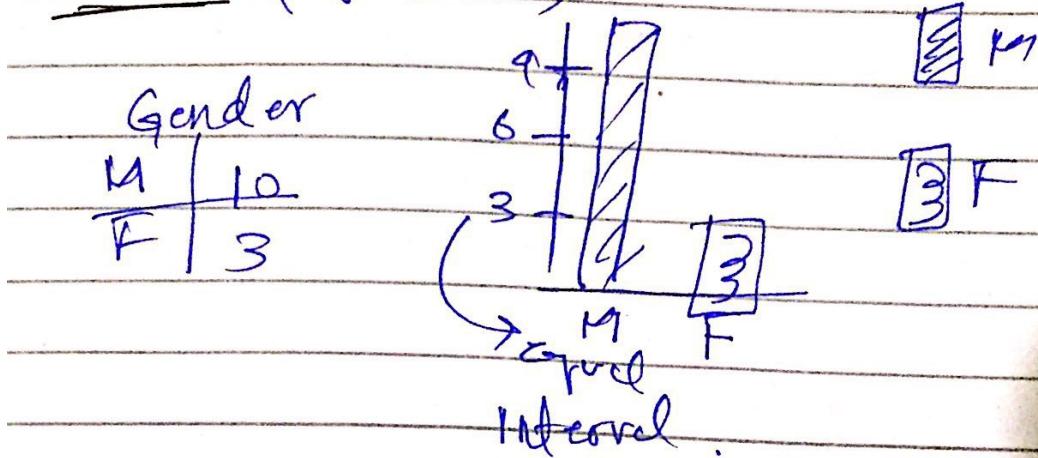
Provinces of Pak ✓

↳ Can do more segmentation.

At our level make sure to divide in  
equal intervals.

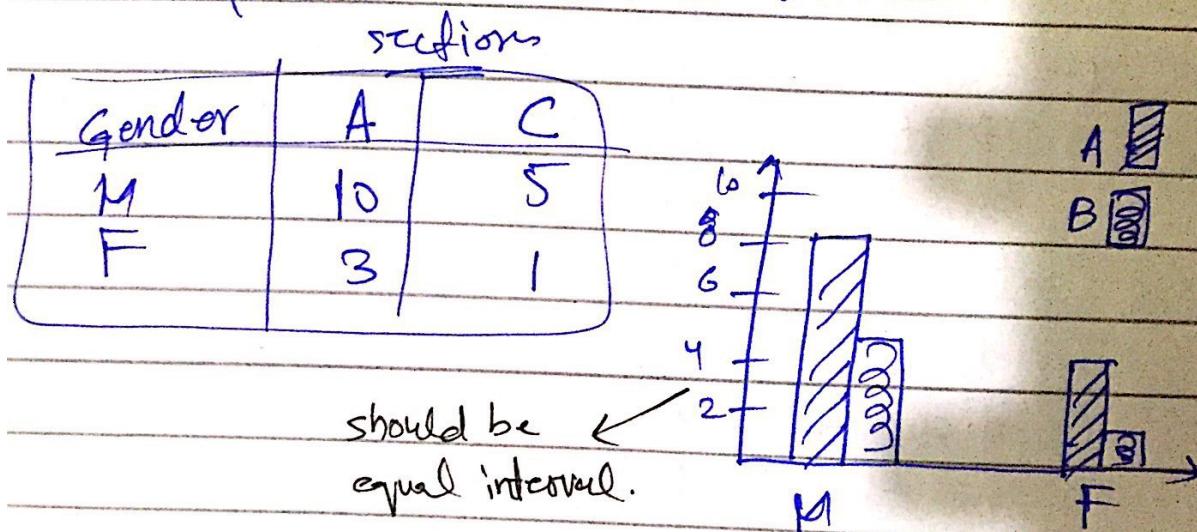
→ Simple Bar chart.

• used for qualitative with single variable. (eg Gender)



→ Multiple Bar chart:

Qualitative with 2 variables.

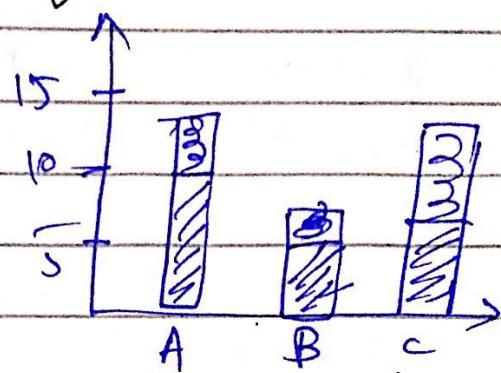


→ Component Bar chart.

Qualitative with 2 variables.

steps -> total size is sum of size of multiple exist  
- if we use sec size X sec or - sec

Gender	A	C	B	
M	80	5	6	21 ? results in ABCD
F	3	1	2	control numbers
	13	8	8	- sum total



drives in  
comp.

مختصر ملکیت اسحاق بوجہا - گردنز  
- جارتی multiples

More vars more dimensions

# Chernoff faces

self study

individual part ← (complex)

of face for specific view

because humans easily recognize

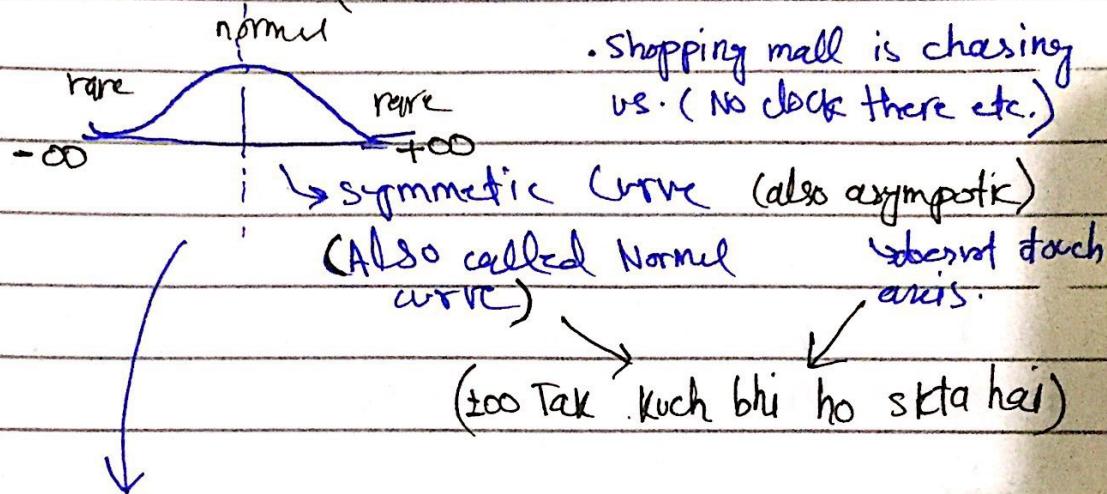
faces.

faces.

→ Histogram.

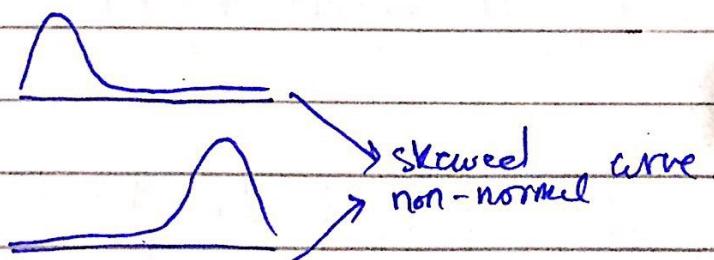
- Quantitative data. (input)

- Normal. (what is normal?)



Normality → things are in a group (cluster)

deviation from normality → abnormality.



→ why we are talking about normality?

↳ normal is a scenario, mostly statistical tools consider data to be normal.

- Normal is easy to understand, abnormality has many cases.

- We are studying descriptive stats till now.

Representative of data  $\rightarrow$  (Min, Max, Avg)

- Measure the center location.

$\bar{X}$  (balancing point)

- Three Methods to measure center.

(1) Mean (A.M) OR Average.

(arithmetic mean)

(2) Median

(3) Mode

$\rightarrow$  Mean.

$$\text{A.M} = \bar{X} = \frac{\sum_{i=1}^n}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Avg is center of data.

- Average is sensitive by extreme observation.

↓  
outlier.

$$\cdot \frac{1+2+3+4+(500 \text{ outlier})}{5}$$

- 3 r. are representative data for large vs 3 r. extreme values vs avg j1

. outliers can be 1 or 2 i.e. if data else you are making data heterogeneous.

. We will talk about real life data.

## (ii) Median

↳ middle value of arranged dataset

steps

- (i) Arrange Data.
- (ii) Find Middlemost data.

even observation



2 middle most

2

$$1, 2, 3, 4, 5, 6 \uparrow \left( \frac{3+4}{2} \right)$$

odd observation



middle val.



1, 2, 3, 4, 5

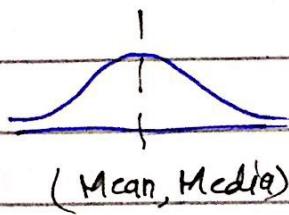
. Median is not affected by extreme observation.

. Median is robust or less - ~~sensitive~~ method

- Mean is balancing point.
- Median is 50% of data.

Mean & Median are only for quantitative data.

because avg of quality is useless.



symmetric curve  
In ↑ mean, median are equal

we will prefer Mean here.

We can use histogram to draw data if it is normal or not.

Assess the shape if symmetric then go for Mean, else go for Median.

Our new grading is based on median to normalize outliers.

(Histogram, Mean, Median) → observe then go for mean, median

Mode: → (Average quality)

↳ no. of observation repeated most often time.

same we can find mode of quantitative but it is not useful.

- Useful for questionnaires.

Most used soap

↓

↳ through mode

LUX 1

Dettol 5

Captivi 6.

avg → 4 is  
useless here

- Weighted A.M.:

$$\bar{X}_w = \frac{\sum w_x}{\sum w}$$

Average is not enough to make decision.

↳ until we know values how much deviate from center.

→ Measure of dispersion:

i) Range

+ not popular or usable.

diff b/w max & min

$$\text{range} = \text{max} - \text{min}$$

max diff between two points.

1  
2  
3  
4  
5

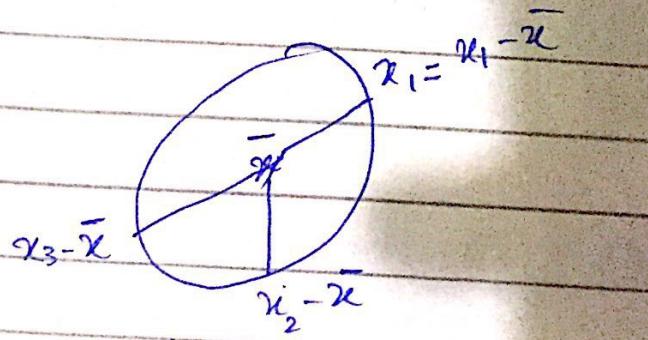
$$\text{range} = 5 - 1 = 4$$

(ii) Avg Dispersion

$$\sum \frac{(x - \bar{x})}{n}$$

$$(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2$$

5



for abm values 12345

Avg disp 0

Sum of the deviation of each observed from their mean = 0.

$N \rightarrow$  size of population

$$\sum \frac{|x - \bar{x}|}{n}$$

mean deviation M.D

question is why we are making the without any mathematical treatment.

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

Variance

unit of measurement  
will get square

$\sqrt{\text{variance}} \Rightarrow$  standard deviation

-  $\bar{x} \rightarrow$  average < centre of values >

. Standard deviation is more accurate  
than mean deviation.

$$\bar{x} \pm S.D$$

Mean in population  
data is labelled  
as  $M(\mu)$

2 → variation  
6 → S.D

## \* 3 Sigma limits:

$5 \pm 2 \rightarrow 3 \leftrightarrow 7$   
 5 is center with 2 variation.  
 more data stay here.

\* in continuous data, interval has more accuracy.  
 4pm? (3:55 - 4:00 pm)

$$\begin{aligned} \text{one sigma interval } \bar{x} \pm 1\text{SD} &= 68.5\% \\ \text{two } \sim \sim \bar{x} \pm 2\text{SD} &= 95.7\% \\ \text{three } \sim \sim \bar{x} \pm 3\text{SD} &= 99.5\% \end{aligned} \quad \left. \begin{array}{l} \text{obs will} \\ \text{lie here} \\ \text{at min.} \end{array} \right\}$$

data = 1, 2, 3, 4, 5

find SD:

$$\frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5} = \frac{4+1+1+2+4}{5}$$

$$\text{variance} = 2 \Rightarrow \text{SD} = \sqrt{2} = 1.41$$

If 2 SD below  $\rightarrow$  fail

" " above  $\rightarrow A$ .

If 3 SD below  $\rightarrow$  full fail

" " above  $\rightarrow A+$

$\therefore SD$  is relative  
to mean

28-09-21

Quartile:  $\rightarrow$  cut data in 4 equal parts.

$Q_1$        $Q_2$        $Q_3$

— + — + —  
25%.      50%.      75%.

$\rightarrow$  Measure of dispersion cont...

$\rightarrow$  Interquartile Range (IQR):

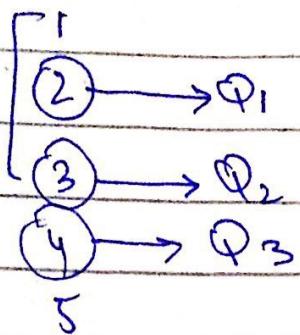
$$IQR = Q_3 - Q_1$$

$$= (\text{Last quartile} - \text{First quartile})$$

IQR is relative to  
median base.

• Quantiles help to know above or how much  
below you are from a reference point

$\hookrightarrow$  percentile in  
NTS



we have 50 obs array 1-25

4 tell 25<sup>th</sup> for median

↳ smart work.

$$IQR = Q_3 - Q_1 = \Delta$$

Median  $\pm IQR$  representation.

$3 \pm 2$

• First preference is  $\bar{x} \pm S.D$

but if outlier is skewing

↳ Median  $\pm IQR$ .

•  $IQR$  is not sensitive to outliers

↳ because end points not included.

•  $3 + \text{accuracy}$  is used in industry for quality assurance

statistical quality control

1 crate 1 egg large scale unit can bear less accuracy.

1 dozen 1 egg small  $\leftarrow$  cannot  $\leftarrow$   $\leftarrow$

→ these all are  
absolute measures

$$\bar{X} \pm S.D$$

$$\text{Median} \pm 1\text{QR}$$

$$\text{Mode} \pm \text{Range}$$

outlier effect

outlier does not effect

cannot contain outlier

⇒ those are qualities.

## Box Plot ./ Wisker Plot .

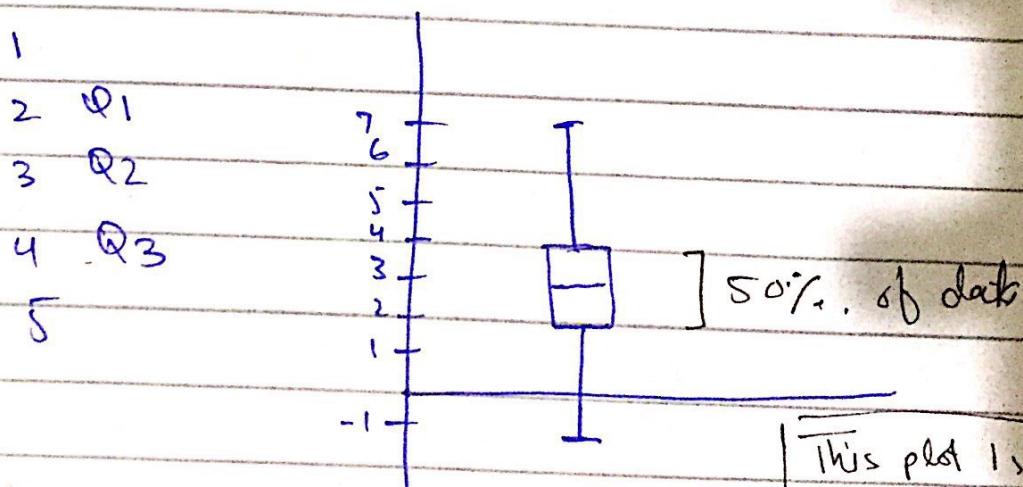
↳ 5 point summary.

: includes:

- ①  $Q_2$  Median
- ②  $Q_1$  lower quartile
- ③  $Q_3$  upper quartile
- ④ if value  $< Q_1 - 1.5\text{IQR}$  or value  $> Q_3 + 1.5\text{IQR}$

then value is an outlier.

• 50



Let an outly = 50 = .

This plot is  
used to identify  
outliers in data  
set

$50 \pm 10 \text{ (SD)}$  → does not mean data vary  
 $3 \pm 1 \text{ (SD)}$  kar raha hai.  
 It only tells how data is scattered from center.

if center diff,  
not comparable.

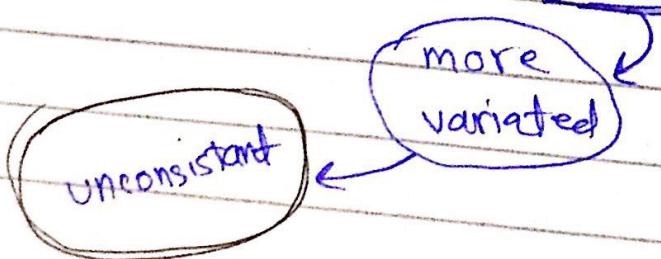
⇒ Coefficient of Variation (C.V)  
 " " S.D.

↳ relative measure of S.D.

. if want to make any thing unit free , divide it with same unit thing.

$$C.V = \frac{S.D}{\bar{X}} * 100 \quad \text{in term of percentage.}$$

	$\bar{X}$	S.D	C.V
D <sub>1</sub>	50	10	20%.
D <sub>2</sub>	3	1	33.33%.



• It's not good to make decision on avg.  
so, we move towards (C.V).

$$= \frac{33.33}{20} = 1.666$$

if ratio same then both same  
else  $1.6665 \rightarrow 0.8665$

. 66.65% data above is more  
variated

if ratio = 2  $\rightarrow$  2 times greater.

→ Coefficient of IQR :

$$C.I.Q.R = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

→ coefficient of Range:

$$C.R = \frac{\text{Max} - \text{Min}}{\text{Max} + \text{Min}}$$

End of one portion of  
descriptive stats

calculating.

- chances of uncertainty,  
Probability

Lec # 9

05-10-21.

"Probability"

In terms of  
numerical  
value.

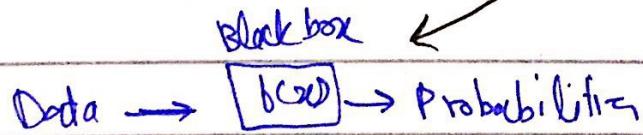
- Quantification of uncertainty is probability.

- Probability → measuring chances of uncertainty.

- Two schools of thoughts / approaches to calculate probability

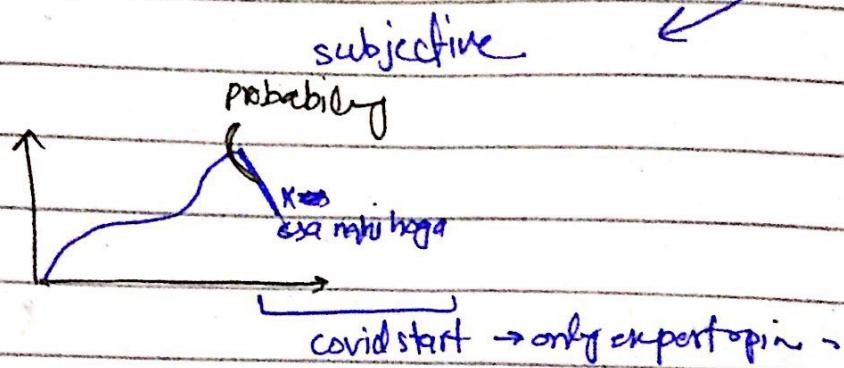
(i) Subjective

(ii) Objective.



\* subjective can be biased  
as it is expert opinion.

• Data + info  $\rightarrow$  decision on experience



• If dataset is very small objective support is not good.

## Experiment

• Any planned activity

## Random Experiment

( $\Rightarrow$ )  $\text{जिसे विकल्प - कहा जाता है}$

outcomes are known but exact outcome not known.  
point ① point ②

planned activity ki outcomes beta hon kew  
exact na beta ho.

If experiment is not random, we cannot apply probability

### \* Sample Space \*

(gather all possible outcome/results of a random experiment then that set or space  $\Rightarrow$  call sample space.)

Samp space of coin = {H, T}.

### \* Outcomes \*

- each element in sample space is outcome

### \* Event \*

- outcomes in which we are interested

e.g. event  $\Rightarrow$  3 < 6 | 1  $\leq$  4 on dice

Trick is not random experiment

In probability we deal with random experiment only.

$$P(A) = \frac{\text{no. of favourable outcomes / event}}{\text{total no of actions}}$$

Theorem 1

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

$n(A)$  no of occurrence A.  
 $n(S)$  no of elements of sample space.

relative frequency based definition of probability.

$$0 \leq P(A) \leq 1$$

always.  
probability is scaled.

Null event  $\rightarrow$  event with 0 probability.  
sure event  $\rightarrow$  event with 1 probability.

chances = probability \* 100

70% chances of rain  
- 30% chance of no rain  $\rightarrow$  0% - 100% are same ultimate large

50% H & 50% Tail

prob ratios  $\rightarrow$  1:1, sample size.

3. If true  $\rightarrow$  3rd converges

on large scale air system ban jata hai.

$$P(S) = 1$$

Probability of sample space = 1

↳ sure event

$\emptyset$   $\rightarrow$  dis  $\leftarrow$  other than sample space.  
null event.

$$\text{A} \cup \bar{A} = S.$$

$$P(A \cup \bar{A}) = P(S) = 1.$$

$$P(A) + P(\bar{A}) = 1.$$

$$P(\bar{A}) = 1 - P(A)$$

Theorem 2

$$A \quad \bar{A}$$

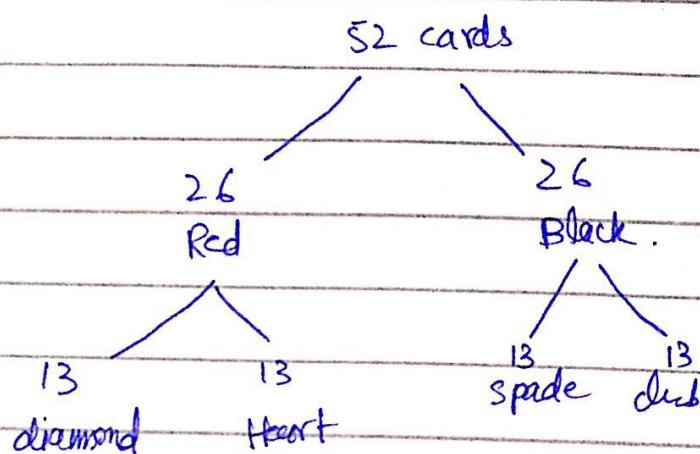
+

1

infinite event mein ik prob pata ho to baki  
sab theorem 2 se nikal sakte hain.

Lec#10

08-10-21



2  
3  
4  
5  
6  
7  
8  
9  
10

J  
Q  
K  
A

) face card

\* Never though probability on  
paper, just write it.

prob of face card

$$P(F) = \frac{n(F)}{n(S)} = \frac{12}{52}$$

$$P(\bar{F}) = 1 - P(F) = 1 - \frac{12}{52} = \frac{40}{52}$$

A, B

P(A or B)

P(A and B)

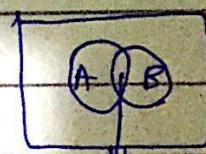
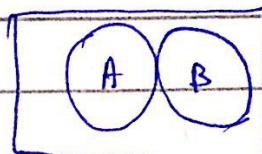
{ OR  $\rightarrow \cup \rightarrow +$   
and  $\rightarrow \cap \rightarrow *$

As probability is in points hence + will increase it while \* will decrease it.

Properties of Events:

(i) Mutually Exclusive Events.

↳ if events cannot occur together

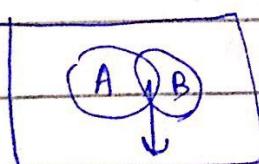


$$A \cap B = \emptyset$$
$$P(A \cap B) = 0$$

null event.

(ii) Mutually inclusive:

↳ if events can occur together



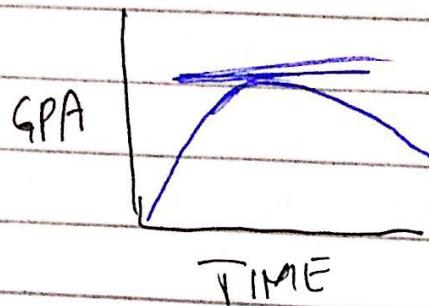
$$A \cap B \neq \emptyset$$

$$P(A \cap B) \neq 0$$

(i) Independent / Dependent Event.

↳ if occurrence of one do not affect other

it is called independent else  
dependent.



• if mutually exclusive we will not check indep  
4 dependent

• if mutually inclusive then we will check  
dependence or independence.

joint probability theorem.

3rd theorem

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$$



↓  
mutually exclusive,

~~P(A and B)~~ Non-Mutually exclusive

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

- Events of common  $\rightarrow$   $\overline{A \cap B}$

→ prob of A or B.

A = six on dice ; B = 1 on dice.

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

$$P(A \cup B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

A = Red card

B = King card

A or B

→ remove double

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

$$= \frac{26}{52} + \frac{4}{52} - \frac{2}{52}$$

$$= \frac{28}{52}$$

$P(A \cap B)$  calculate.

↳ Independent.

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

• Dependent .

• Conditional Probability

• Baye's Theorem.

• most widely used theorem in CS.

# exercise chapter 2.

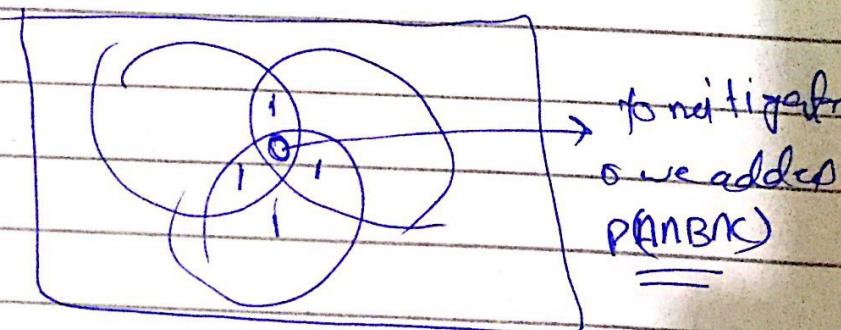


→ if we have  $k$  events

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k)$$

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i)$$

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C). \end{aligned}$$



Remember to adjust doubly

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_k) &= \sum_{i=1}^k P(A_i) - \sum_{\substack{i=1 \\ i \neq j}}^k \sum_{j=1}^k (P(A_i \cap A_j)) + \\ &\quad \dots \quad \text{dependent/odd} \\ &\quad \sum_{\substack{i=1 \\ i \neq j \\ i \neq k}}^k \sum_{j=1}^k \sum_{k=1}^k (P(A_i \cap A_j \cap A_k)) \dots \quad \text{+} \\ &\quad \sum_{\substack{i=1 \\ i \neq j \\ i \neq k \\ \dots}}^k \sum_{j=1}^k \dots \sum_{k=1}^k (P(A_i \cap A_j \cap \dots \cap A_k)). \end{aligned}$$

## Possibility vs Probability.

no. of outcomes

e.g. H, T in  
coin toss.

↳ favorable  
total

- Possibilities are in number.
- Probability is like occurrence.

Two dice rolls

	1	2	3	4	5	6
1	11	12	13	14	15	16
2	21	22	23	24	25	26
3	31				36	
4	41				46	
5	51				56	
6	61	62	63	64	65	66

→ Counting technique.

• principle of counting:

event A has m result & B has n

combined result = m × n.

\* don't confuse probability & possibility.

3 letters & 4 digits  $\rightarrow$  no place

$$\underline{26} \underline{26} \underline{26} \quad \underline{10} \underline{10} \underline{10} \underline{10}$$

$$\text{result} = \underline{26 \times 26 \times 26} + \underline{10 \times 10 \times 10 \times 10}$$

ATM PIN 4 digits

10 10 10 10  $\rightarrow$  10,000 possibilities  
only.

Phone S14.

$$0300 - \underline{10} \underline{10} \underline{10} \underline{10} \underline{10} \underline{10} = 10^7$$

1  
2  
3  
4  
5  
6  
7  
8  
9

$\rightarrow$   $\begin{array}{c} 8 \\ 10 \end{array}$  mark over

Complete.

same ( $n \times n \times n \times n$ )

unequal (nam)

\* dice, if one 6 then 6 can appear again.

Precedingly Reduce.

\* Repetition is not allowed.

• cards

• concept of factorial.

!

• one by one reduce ho raha hoi.

→ Factorial:

$$n! = n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1$$

• objects w/ their arrangements.

-  $3! = 3 \times 2 \times 1$ , i.e.,  $3! = 6$ .

Permutations  
order do matter  
 $AB \neq BA$

order don't matter  
 $AB = BA$

Combinations

A B C  
obj 3, arg=2

AB BA

BC CB

AC CA

$$3P_2 = 6$$

A B C

AB = BA

BC = CB

AC = CA

$$BC_2 = 3$$

$nPr$ ,  $nCr$  permit  $r \leq n$  order  $\uparrow$

- combinations

$$nPr = \frac{n!}{(n-r)!}$$

$$nCr = \frac{n!}{(n-r)!r!}$$

$$\text{Ex} \quad \frac{10 \times 9 \times 8 \times 7 \times 6!}{6!} = \frac{10!}{6!} = \frac{10!}{(10-6)!}$$

to remove  
dup  
arrangments.

$n!$  is special case of ~~factorial~~ permutation.

$$n!P_n = \frac{n!}{(n-n)!} = n!$$

$$nCn = 1$$

Practice: