Lec#1 07-09-2021 Tuesday.

(MT-2005)

- Real life scenarios are based on decision making.

- Data - Info.
- There is no existence of anything without observation.

- Even eye blinking is based on data.

- Environmental & Genetic learning.

→ Statistics:-
Collection of data, presentation of data, analysis of data & decision about the data.

- statistic is umbrella & prob is part of it.

- Deterministic Approach
- Probabilistic Approach

Lec #2 10-09-2021

Computational statistics
↳ Data Mining or i like fields =

(estimation, prediction, forecasting)
All backed by statistics.

- self training algorithms (global prediction)
- Probability theory in prediction

→ Population:

totality of anything

- population study.

- Problems in population study:
  - Time
  - Cost (can be funded by UN to NGO govt.)
  - We cannot do population testing. (unit destroy)
    ↳ (can't check each candy etc)

- Pakistan Bureau of Statistics controlls census.

- Due to issues in population we follow Sample Method.

- Sample: → homogenous heterogenous (if so then make strata).
  ↳ (part of population) which contains characteristics of population.

(i) Are you giving valid data?
(ii) Are proper conventions being followed to get data.

- sample should be propotion to size of population.

- Sometime people give wrong info that causes wrong ~~inf~~ decision.

UCI Machine Learning repo

- Data is the truth.
  we have wrong perceptions.

→ Parameters:
  ↳ set of characteristics.

- Books

(i) Probability & Statistics for engr. & scientist.
(walpole & Myer)

(ii) A first course in probability (Ross Sclolom)

- Collection of Data -

primary data
- First handed data or raw data.
(خام یا براہ راست)
(questinare, discussion etc)

secondry data.
- second Hand data or statistical tool applied on primary data creates secondry data.
(Nadra etc)

State bank's web contains lots of data.
Machine learning data sets

Lec#3 14-09-21

- 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 fidis 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:
      - Class must be from 6,7. → expert opinion.
        - If classes > 7 make size of classes accordingly.
      - class size can be fix or not.
      - Take that class size jis se no. of class 6-7 ho.

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

classes (Age)	C.F	Frequencies.	%
0-10	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	

(open end / close end) ← 0-10

must. → total data = sum

- Extreme observations can also impact data.

0-9
10-19
20-29

✗ wrong approach
disturbs continuity

15 se kam wala:

C.f → comulative frequency.

sum += previous frequency.

↳ it can be rerse order.

↳ 15 se zaya wala.

→ F.d for qualitative data:

Gender	
Male	15
Female	5
	20

→ We cant make classes here like 1-2 ? is useless. We have to represent it in categories.

If categories exceed size, so we divide it.

cities of pak ✗

provincies of Pak ✓

↳ can do more segmentation.

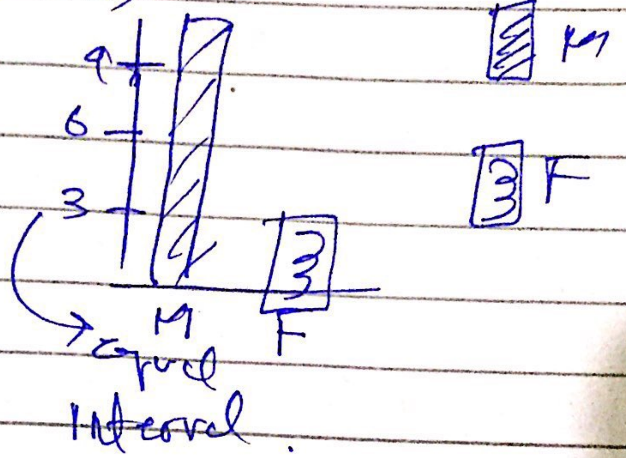
. At our level make sure to divide in equal intervals.

→ Simple Bar chart.

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

Gender

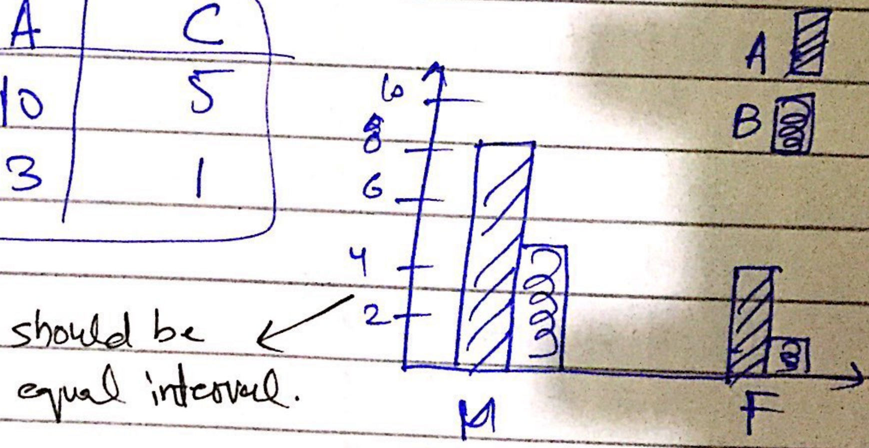
M	10
F	3



→ Mutiple Barchart:

Qualitative with 2 variables.

	sections	
Gender	A	C
M	10	5
F	3	1



→ Component Bar Chart.

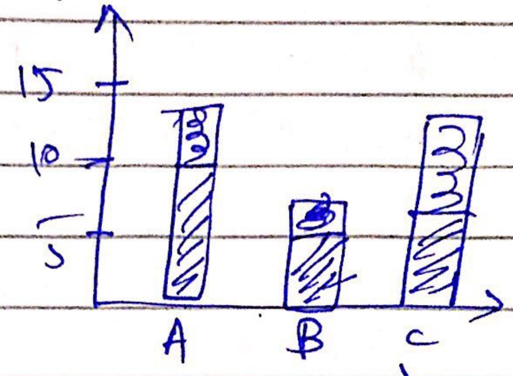
Qualitative with 2 variables.

multiple کو ہم ایسے بھی بنا سکتے ہیں اگر total لے کر secA کو [illegible]

[illegible] سب sec X کے بجائے sec کے [illegible] exist [illegible]

Gender	A	C	B
M	10	5	6
F	3	1	2
	13 ✓	6 ✓	8 ✓

کیا ABC کے علاوہ بھی ہے ؟ 21
یہ پتہ نہیں ہوتا کی وجہ سے
total نہیں ہے ۔



→ drives in comp.

صرف ٹوٹل کے لیے یہ استعمال ہوگا ۔ ورنہ
multiple چارٹ ۔

More vars more dimensions

chernooff faces — self study (complex)

→ individual part of face for specific var because humans easily recognize faces.

Bar charts only for qualitative

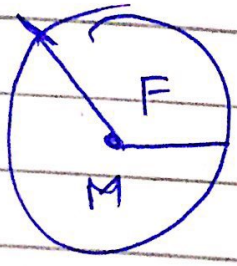
- Pie Chart → (one variable qualitative)
  - ↳ only feasible for data passed for simple bar chart.

- Alternative of simple bar chart.
- convert data in angle.

Gender	
M	10
F	5
	15

$10/15 \times 360 = 240°$

$5/15 \times 360° = 120°$



(use D for angles)

IBM → data analysts.

SPSS most famous.

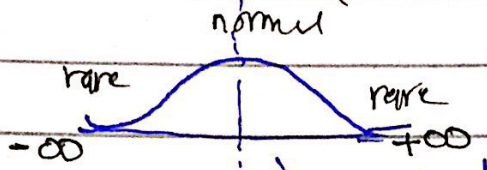
Lec#4

17-09-21

→ Histogram.

- Quantitative data. (input)

- Normal. (what is normal?)



- Shopping mall is chasing us. (No clock there etc.)

↳ symmetric curve (also asymptotic)

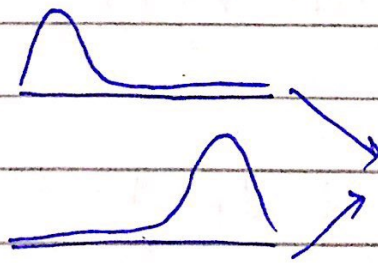
(Also called Normal curve)

doesn't touch axis.

(±∞ Tak kuch bhi ho skta hai)

Normality → things are on avg (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.

Lec#5

21-09-21

- We are studying descriptive stats till now.

Representative of data → (Min, Max, Avg)

- Measure the center location.
  - ↳ (balancing point)
- Three Methods to measure center.
  - (1) Mean (A.M) OR Average.
    - (or arithmetic mean)
  - (2) Median
  - (3) Mode

→ Mean.

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

Avg is center of data.

- Average is sensitive by extreme observation.
  - ↓
  - outlier.
- $\frac{1+2+3+4+(500)}{5}$ outlier

اگر avg میں extreme value ہے تو وہ avg پورے data کا representative نہیں ہوتی۔

. outliers can be 1% or 2% of data else you are making data heterogenous.

. We will talk about real life data.

(ii) Median

↳ middle value of arranged dataset.

steps

(i) Arrange Data.

(ii) Find Middle most data.

even observation

↓

$\frac{2 \text{ middle most}}{2}$

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

odd observation

↓

middle value

1, 2, (3), 4, 5

. Median is not affected by extreme observation.

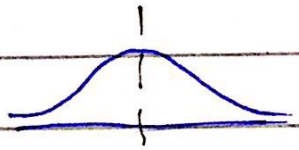
. Median is robust or ~~un~~ less sensitive method.

Scanned with CamScanner

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

Mean & Median are only for quantitative data.

↳ because avg of quality is useless.



(Mean, Media)

In ↑ symmetric curve mean, median are equal

we will prefer Mean here.

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

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

- Our new grading is based on median to normalize outliers.

- (Histogram, Mean, Media) → observe then go for mean, medians

→ qualitative data

Mode: → (Average quality)

↳ no. of observation repeated most often time.

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

- Use ful for questimaire.

Most used soap
↓
6 throw mode

LUX 1
Detol 5
capri 6.

avg → 4 is useless here

- **Weighted A.M:**

$$\bar{X}_w = \frac{\sum WX}{\sum W}$$

CGPA	W	WX
2·1	2	4·2
2·8	3	8·4
3·8	3	11·4
2·9	4	11·6

$$\bar{X}_w = 2.966 = \frac{\Sigma WX}{\Sigma W}$$

• Weighted A.M is simple A.M with weights.

• If weights are equal then WAM will become simple A.M.

Q. Write a Note on Quantials

cut data on multiple points.

(i) def
(ii) formula
(iii) Applic~
(iv) Importance.

quartile
decile
percentile

Acem these.

Lec # 6

27-09-21

- Average is not enough to make decision.
  ↳ untill we know values how much deviate from center.

→ Measure of dispertion:

i) - Range + not popular or usable.

diff btwn max & min

range = max - min.

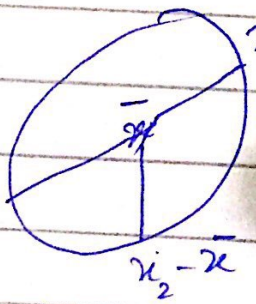
max diff between two points.

1
2
3
4
5

range = 5 - 1 = 4.

(ii) Avg Dispertion

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



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

∵ for above values 1 2 3 4 5

Avg disp 0

sum of the deviation of each observ from their mean = 0.

N → size of population

$$\sum \frac{|x-\bar{x}|}{n} \rightarrow \text{mean deviation (M.D)}$$

question is why we are making +ve without any mathematical treatment.

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

- unit of measurement will get square

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

ہر value اپنے centre سے avg کتنا دور ہے ۔

- Standard deviation is more accurate then mean deviation.

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

Mean in population data is labled as μ (mu)

$\sigma^2 \rightarrow$ variation

$\sigma \rightarrow S.D$

* 3 Sigma limits:

$5 \pm 2 \longrightarrow 3 \longleftrightarrow 7$

5 is center with 2 variation.

max data lies here.

* in continus data, interval has more accuracy.

4pm ? (3:55 - 4:00 pm)

one sigma interval $\bar{X} \pm 1SD = 68.5\%$

two " " $\bar{X} \pm 2SD = 95.7\%$

three " " $\bar{X} \pm 3SD = 99.5\%$

} obs will lie here at min.

data = 1,2,3,4,5

find SD.

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

variance = 2 $\Rightarrow$ S.D = $\sqrt{2}$ = 1.41

If 2 SD below → fail

" " above → A.

If 3 SD below → full fail

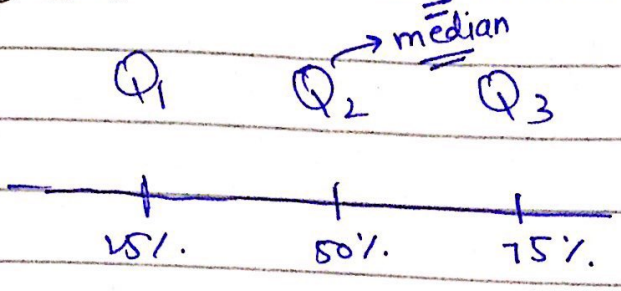
" " above → A+

Lec# 7

28-09-21

∵ SD is realtive to mean

Quartile: → cut data in 4 eq parts.



→ Measure of dispersion cont...

→ Interquartile Range (IQR):

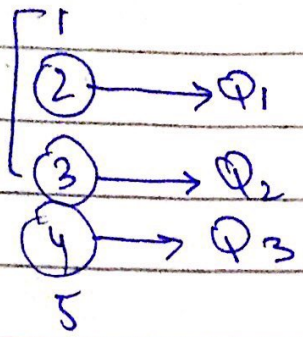
$$IQR = Q_3 - Q_1$$

= (last quartile - First quartile)

IQR is relative to median base.

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

↳ percentile in NTS



we have 50 obs array 1–25
4 tell 25th for median
↳ smart work.

$$IQR = Q_3 - Q_1 = 2$$

Median ± IQR representation.

3 ± 2

- First preference is $\bar{X} \pm$ S.D
  but if outlier is skewing
  ↳ Median ± IQR.

- IQR is not sensitive to outliers
  ↳ because end points not included.

- 3 + accuracy is used in industry for quality assurance

Statistical quality control

1 crate 1egg large scale unit can bear less accuracy.
1 Dozen 1 egg small " " cannot " " "

these all are absolute measures.

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

Median $\pm$ IQR

Mode $\pm$ Range

outlier effect

outlier doesn't effect

. cannot contain outlier

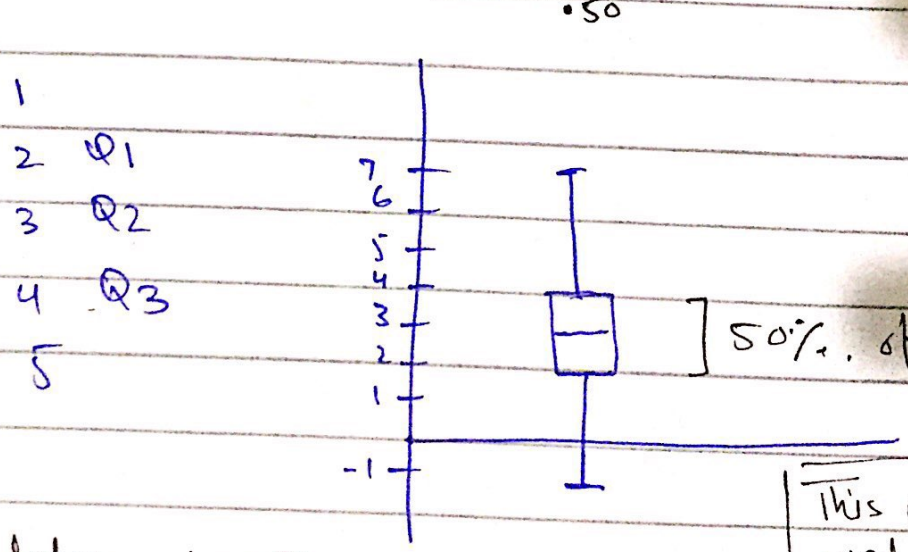
⇒ these are qualities.

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 < $Q_1 - 1.5$IQR or value > $Q_3 + 1.5$IQR then value is an outlier.



Retangle only = SD = .

This plot is used to identify outliers in data set

Lec #8 01-10-21

50 ± 10 (SD) → doesn't means data zagrafiyary kar raha hai. It only tells how data is scattered from center.

3 ± 1 (SD)

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.

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

	$\bar{x}$	S.D	C V
D_1	50	10	20%.
D_2	3	1	33.33%.

more variated

unconsistant

• 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 → 0.6665

• 66.65% data above is more variated

if ratio = 2 → 2 times zayada.

→ Coefficient of IQR:

$$C \cdot IQR = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

→ coefficient of Rang:

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

End of one portion of descriptive stats

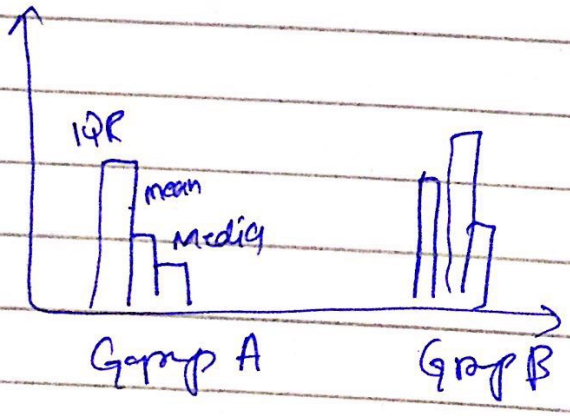
Mix Nature Analysis

Qaulitive vise Quantitative Analysis

demographic variables.

		$\bar{x}$ (SD)	Median (IQR
Male	Rural / urban		
Female	Rural / urban		

- we can also present these value in graph too



(mix nature analysis)

BMI → Body Mass Index

- Uncertainty. ✓
- Probability. (will read)

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 school of thoughts / approaches to calculate probability

(i) subjective (ii) objective.

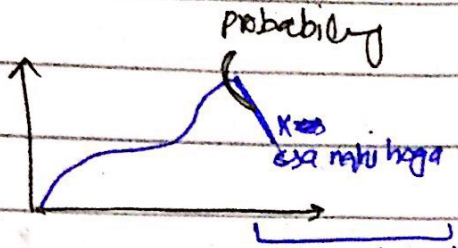
Black box

Data → [box] → Probabilities

* subjective can be biased as it is expert opinion.

- Data + info → decision on experience

subjective



covid start → only expert opin.

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

Experiment

- Any planned activity.

Random Experiment.

(Ex) پتہ ہو کیا کیا آ سکتا ہے - بتا کیا آئے گا یہ اس پتہ

outcomes are known but exact outcome not known.

point (1) point (2)

planned activity ki outcomes pata hon par exact na pata 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 is call sample space.)

samp space of coin = {H, T}.

* Outcomes *

- each element in sample space is outcome

* Event *

- outcomes in which we are interested

جیسے dice میں 4 چاہیے اگر آ جائے تو وہ event ہوگا۔

Trick is not random experiment

In probability we deal with random experiment only.

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

theorem 1

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

$n(A)$ 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 → event with 0 probability.
sure event → event with 1 probability.

chances = probability * 100

70% chances of rain ?

مطلب same state days میں 100 دن ہوں تو 70 دن بارش ہوگی۔

50% H & 50% Tail

جیسے جیسے sample بڑھا تو یہ ratio → probab میں converge ہوگی اور یہ باقی true بن جائے گی۔

. on large scale aik system ban jata hai.

$P(S) = 1$

probability of sample space = 1
↳ sure event

else → other than sample space.
↓
NULL event.

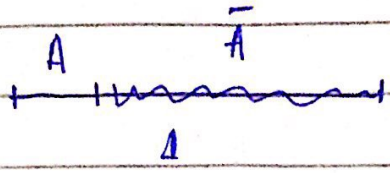
$A \cup \bar{A} = S.$

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

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

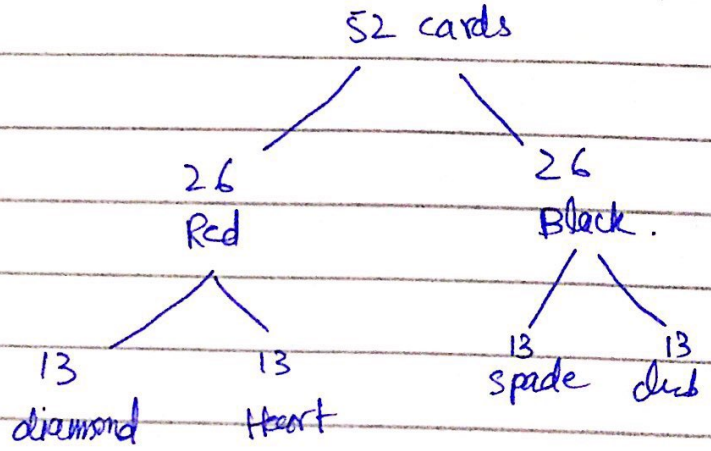
$$P(\bar{A}) = 1 - P(A)$$

theorem 2



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

Lec# 10 08-10-21



2
3
4
5
6
7
8
9
10
J
Q
K
A
} face card

* Never through propbability 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)

(union)

or → ∪ → +

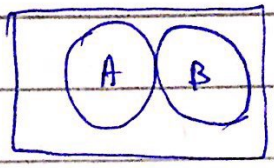
and → ∩ → *

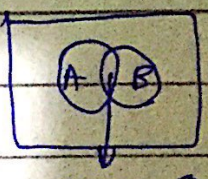
A-> probability is in points hence + will increase it while * will decrease it.

Properties of Events:

(i) Mutually Exclusive Events.

↳ if events cannot occur together





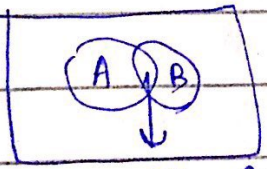
$A \cap B = \phi$

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

null event.

(ii) Mutually inclusive:

↳ if events can occur together.



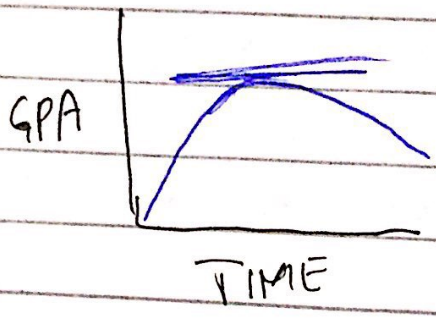
$A \cap B \neq \phi$

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

head/tail — picking cards from stack.

(i) Independent / Dependent Event.

↳ if occurance of one donot affect other it is called independent else dependent.

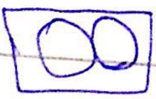


- if mutually exclusive we will not check indep & dependent

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

~~joint probability theorem.~~ **3rd theorems**

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



↓ mutually exclusive,

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

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

ک ایک دفعہ common نکال دو گے،

→ Prob of A or B.

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

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

$P(A \cup B) = 1/6 + 1/6 = 2/6 = 1/3$

A = Red card
B = King card.

A or B → remove doubles

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

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

$= \frac{28}{52}$

$P(A \cap B)$ calculate.

→ . Independent.

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

. Dependent.

. Conditional Probability

. Baye's Theorem.

. most widely used theorem in CS.

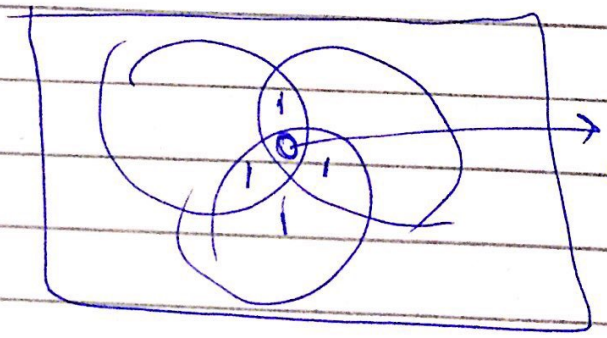
exercise
chapter 2.

→ if we have k events

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

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

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



→ to mitigate we added $P(A \cap B \cap C)$

Remember to adjust doubling

$$P(A_1 \cup A_2 \cup \dots A_k) = \sum_{i=1}^{k} P(A_i) - \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} (P A_i \cap A_j) +$$

depend even/odd

$\Sigma \Sigma \Sigma (P A_i \cap A_j \cap A_k) \dots \dots \dots$ (+)

$\Sigma \Sigma \dots \Sigma P(A_i \cap A_j \dots A_k)$.

A first course in prob & stats
Ross Seldom.

Lec #11

12-10-21

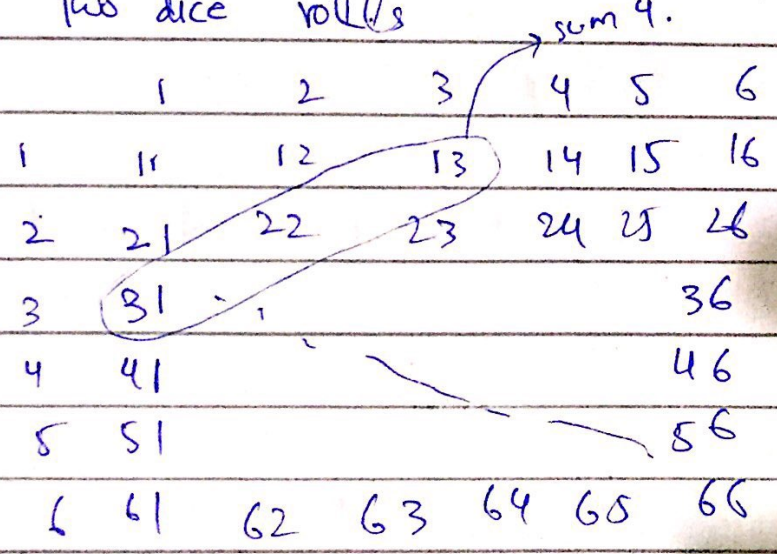
Possibility vs Probability.

no of outcomes
e.g. H, T in coin toss.

↳ favorable / total

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

Two dice rolls



→ Counting technique.

- principle of counting.

event A has m result & B has n
combined result = m×n.

*donot confuse probability & possibility.

3 letters & 4 digits ⟶ no plate

26 26 26 10 10 10 10

result = 26×26×26×10×10×10×10

ATM PIN 4 digits

10 10 10 10 → 10,000 possibilities only.

Phone SIM.

0300 - 10 10 10 10 10 10 10 = 10^7

1
2
3
4
5
6
7
8
9 ⟶ (8 / 10) = max ven

Complete.	Preceding by Reduce.
	* Repetition is not alloed.
same (n x n x n x n)	• cards. • concept of factorial.
unequal (n x m)	
* dice, if one 6 then 6 can appear again.	• one by one reduce ho raha hai.

→ **Factorial:**

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

• objects n their arrangments.

جو چیز ایک بار استعمال ہو جائے دوبارہ نہیں ہوگی۔

Permutations (→ oder do matter, $AB \neq BA$)	Combinations. (→ order donot matter $AB = BA$)
A B C	A B C
obj 3, arg = 2	AB = BA
AB BA	BC = CB
BC CB	AC = CA
AC CA	$^3C_2 = 3$
$^3P_2 = 6$	

Scanned with CamScanner

اردو میں permutation کے لیے ترتیب اور combination کے لیے انتخاب کا لفظ استعمال کرتے ہیں۔

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

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

$r!$ to remove dup arrangements.

$$^{10}P_7 = \frac{10\times9\times8\times7\times6!}{6!} = \frac{10!}{6!} = \frac{10!}{(10-7)!}$$

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

$$^nP_n = \frac{n!}{(n-n)!} = n!$$

$$^nC_n = 1$$

Practice: