↳ Statistics ⇒ Method of collecting, organising and analysing data in a such a way that meaningful conclusion can be drawn.

Why ?

ML ↓ Learn Patterns

data → fact | information. that can be stored and measured.

Example → Scores made by Virat Kohli in last five matches

↳ [92, 80, 40, 70, 60]

Ex.2. Height of students in the classroom

↳ [150, 160, 180cm, 180, 190cm]

Motivation behind statistics

① Weather forecast — 45°C
② Sports analysis — SR - VK.
③ Election campaign → Exit Polls → BJP
④ FMCG | e-commerce — DMART
⑤ medical | genetics

Types of statistics

Descriptive → Describe → It consists of summarising and organising data.

Inferential → Inference → Using sample data from a population

→ Complete population.

but can also be calculated for sample.

and calculate statistics.

Sonu → Razi

CEO → Airline industry (Akasa)

→ Descriptive

3 - chijon

① Avg delay of flight
② Sales
③ # of tickets booked.

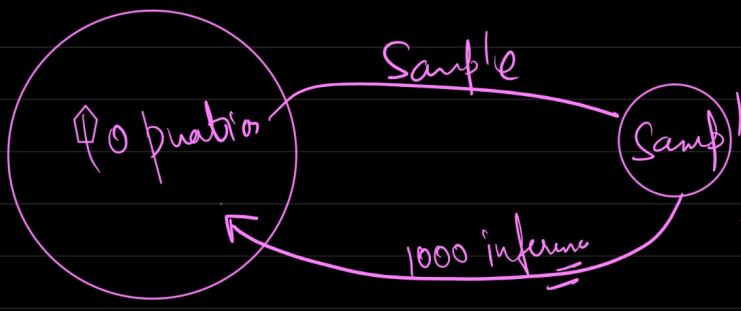
Sakshi

↳ travelling → Kaziranga National park

No of trees

inferential $1000m^2$ → 1000

$100000m^2$ - 100000 trees



Statistics ↳ 1000

140cr

Dipam

Khalid = 80kg ↳ 180kg

Wt.

inferential stats

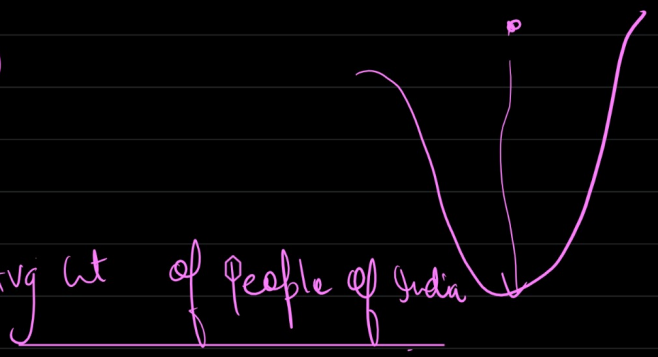
* Class:

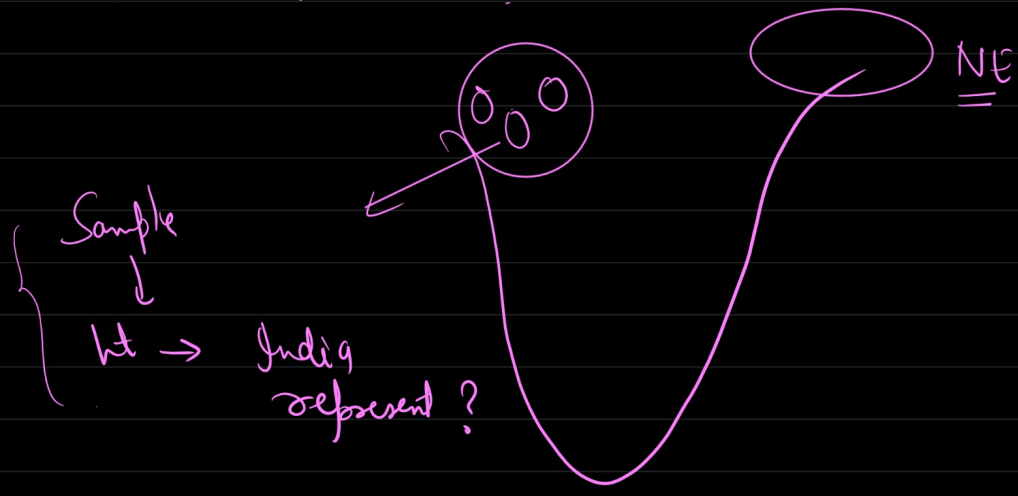
Descri — Avg ht ?
— most frequent ht ?
— Strike rate of VK ?

Descriptive

① Measure of CT ② Measures of Symmetry measure of dispersion

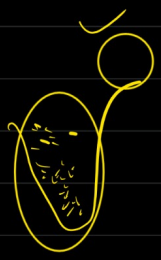
* Inferential





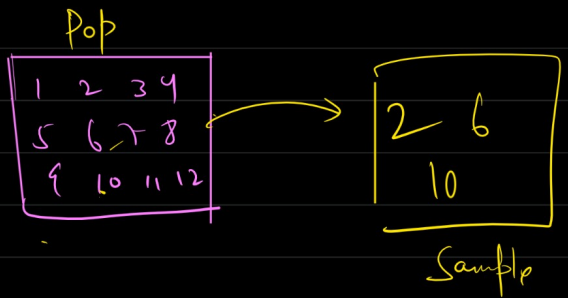
Types of Sampling

① Simple random Sampling.



Each member of pop. has an equal chance of being selection

⇒ 1/n.



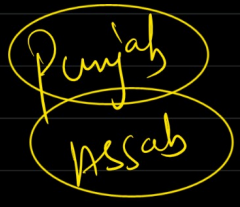
② Stratified sampling

⇓

Strata / layer / groups

③ Cluster Sampling

→ divides the population into groups / clusters. Some of these clusters are randomly selected.



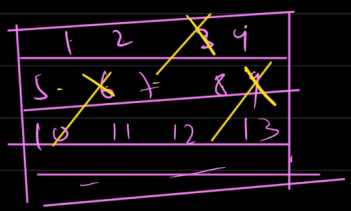
Covid days Covidindia.org

> 100000 — yellow

< 1000 — ( )

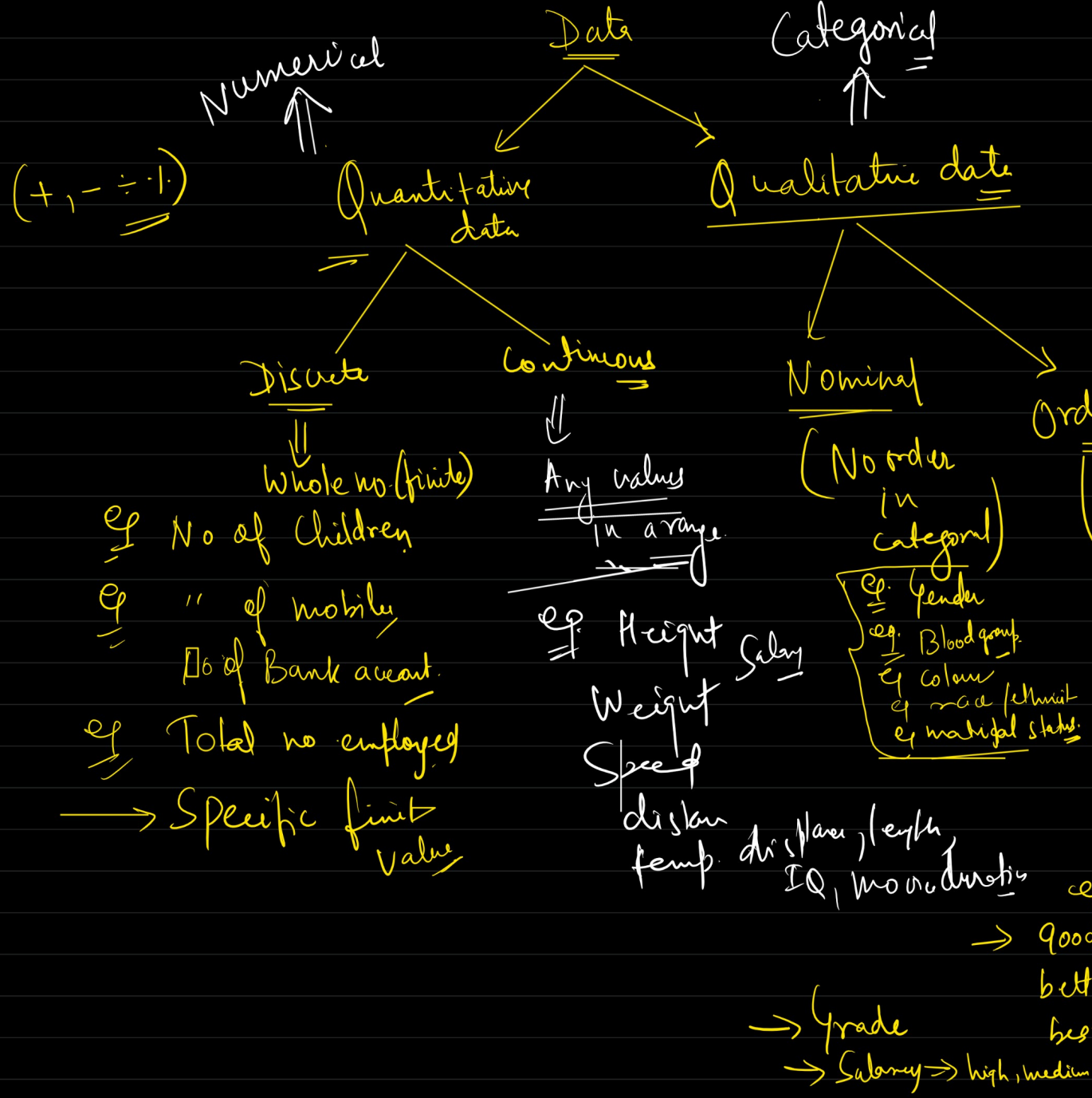
⑨ Systematic Sampling

Every n^{th} element.



* Quota Sampling
* Min Max Sampling
* Convenience Sampling
* Accidental Sampling.

* Types of data (Calculation)



Scales of measurement

→ Military ranks.

- Scales
  - Nominal scale data
  - ordinal scale data
  - interval scale
  - Ratio Scale

① Nominal scale data

→ Qualitative / categorical data

→ Gender, ethnicity, colour, location.

→ No order in the data.

Example

Employees

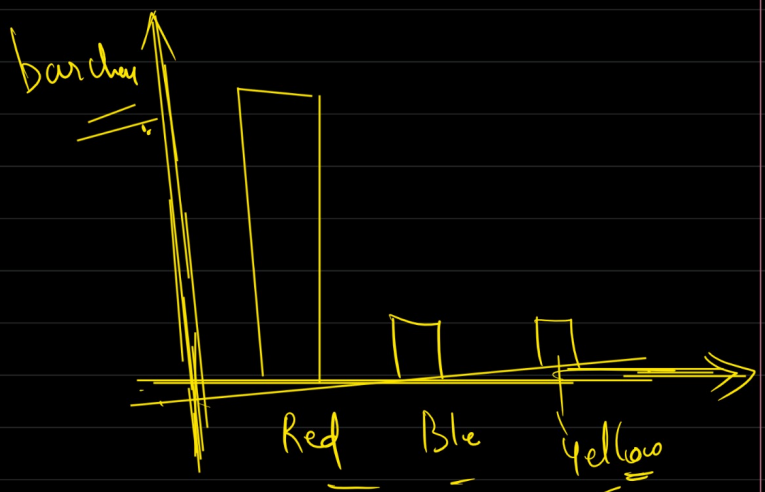
M
F
M
F
M
M
M
M

M - 6
F - 2

* Count.

Red - 5 → 50%
Blue - 3 → 30%
Orag - 2 — 20%

Pie-chart



② Ordinal Scale data

→ Order / rank matters.
→ Difference can not be measured.

Salary	Qualification
H	PG
L	UG
M	12th
H	10th
H	G
L	PG
L	PhD.
M	

Performance of Student

1st
2nd
3rd
4th
:

Tushar 1st
Divyansh 2nd

③ Interval Scale data

→ rank & order has a meaning
→ Difference can be measured (excluding ratio)
→ It doesn't have 0 starting value.

eg. marks
length
Wt
temp
revenue of Comp
ht
Stock Price

90
80

④ Ratio Scaled data

→ Order / rank has meaning
→ Difference and ratio is measurable.
→ It does have a 0 starting point.

wt, ht., Age, time

temperature X

$\frac{\text{Pant}}{\text{Shirt}} = \frac{80}{40} = 1:2$

−5, 0

Temp.

inside 30°c

outside − 60°c

−5 −10 −20

$\frac{30}{60} = 1:2$

Temp can be negative

Data	Nominal	Ordinal	Interval	Ratio
labelled	✓	✓	✓	✓
Order	✗	✓	✓	✓
differ	✗	✗	✓	✓
True Zero Starting Point	✗	✗	✗	✓
Example →	Gender, Religion, Post office code, location	Satisfaction Rating, Grade, Ranks, Military rank	IQ, Temp, Shoe size, Int wt	Ht, Wt, time, Age

* Descriptive Statistics (Summarization of data)

① Measure of central tendency
② Measure of dispersion
③ " " Symmetricity.

* M C T

Central ⇒ What is that one value around which all the data is revolving.

1, 2, 3, 4, 5 — ③

CT → represents the center point of a data.

① mean
② Median
③ mode
} EDA and feature Engineering.

① Mean

{1, 2, 3, 4, 5}

$\frac{1+2+3+4+5}{5}$

$x = 1, 2, 3, 4, 5$

$\sum \rightsquigarrow 1+2+3+4+5$

$\mu = \frac{\sum x_i}{N}$

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

$x = 1, 2, 3, 4, 5$

$\prod x_i$

$1 \times 2 \times 3 \times 4 \times 5$

$x = (8, 5, 6, 7)$

$\sum_{x=0} = 8+5+6+7$

$mean = \frac{\sum_{i=1}^{n} x_i}{n}$

1, 2, 3, 4, 5

$\rightarrow \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$

$= \frac{1+2+3+4+5}{5}$

② Median (Physical mid Point)

4, 5, 2, 3, 1, 2

* Sort the data → 1, 2, 2, 3, 4, 5.

Count the no of elemt = 6

if count is even → avg of two middle elemt.

1, 2, 2, 3, 4, 5

$\frac{2+3}{2} = 2.5$

4, 5, 2, 3, 1

Sort = 1, 2, 3, 4, 5

Count = 5 → odd

median = 1, 2, 3, 4, 5 → 3

.mean() , .median()

Sort ⇒ 1, 1, 2, 3, 4, 5, 6, 7, 7, 8, 9, 9

avg = $\frac{5+6}{2} = 5.5$

eg 1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 9

→ 5 is mean

Scenario-1

1, 2, 3, 4, 5

mean = 3

→ 1, 2, 3, 4, 1000

↳ outlier.

mean = $\frac{1+2+3+4+1000}{5} = \frac{1010}{5} = 202$

Scenario-2

1, 2, 3, 4, 5

Median = 3

Sorting → 1, 2, 3, 4, 1000

Median = 3

* The mean is affected by outliers whereas median is not affected by outliers.

③ mode ⟹ Categorial data

↳ maximum frequency.

{1, 1, 1, 2, 2, 3, 3, 1, 1, 2, 3}

Mode → 1

Col

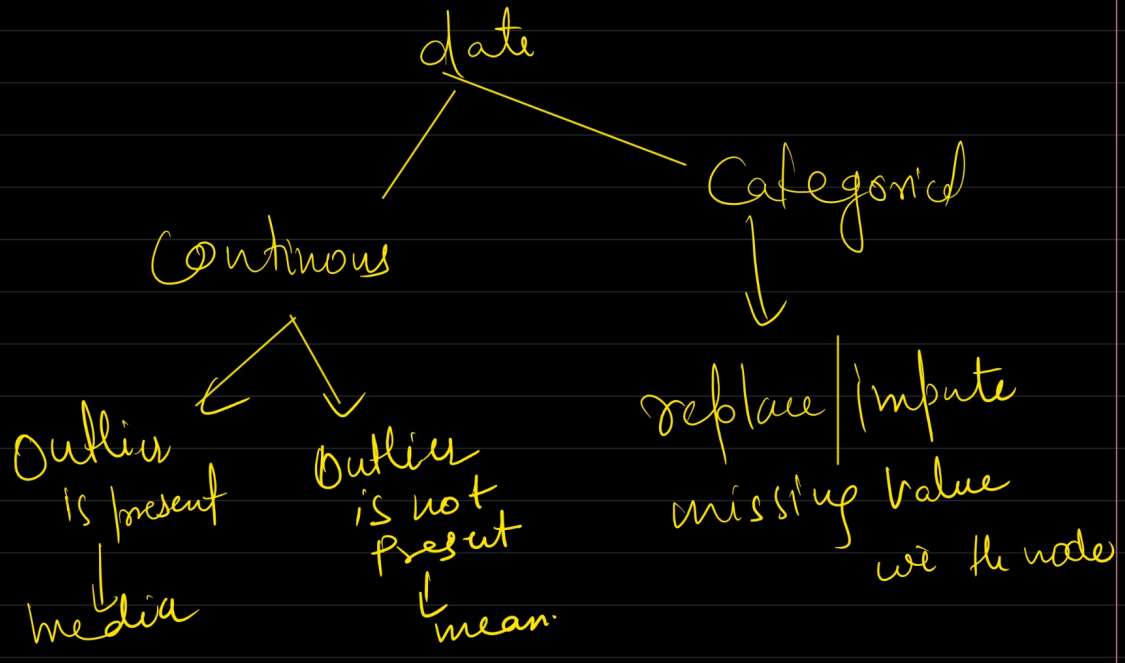
good – 1
better – 2
best – 3

* Use of Central tendency | Numerical → mean, median / Categorial → mode → if outlier present median

Age	Gender	weight	Salary(k)
25	M-1	80	—
26	male Mode = Highest freq.	70	50
—	M-1	30	60
23	M1	—	70
25	F0	60	45

Age → Continuous num dat

mean $\frac{25+26+23+25}{4}$



* Measures of dispersion/Spread

$S_1 = 1, 2, 3, 4, 5$, $S_2 = 3, 3, 3, 3$

mean/median, 3, 3. 3, 3.

* Measure of spread?

- → Range
- → Percentage & Percentile
- → Quartiles (Boxplot)
- → Variance
- → Std dev^n

① **Range** → difference b/w max and min value

$\{1, 2, 3, 4, 5\}$

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

Range is affected by Outlier

$\{1, 2, 3, 4, 5, 1000\}$

$1000 - 1 = 999$

↓

Inter Quartile Range

* **Percentage**

$\{1, 2, 3, 4, 5\}$

What is the percentage of Nos that are odd.

$$\frac{3}{5} \times 100 = 60\%$$

* **Percentile**: A percentile is a value below which a certain percentage of observation lie.

98.5%

Abir = 99%

Abir = 99% लोगो से अधिक score किया है।

$1, 2, 3, 4, 4, 6, 7, 7, 8, 10$

What is percentile rank of 3 ??

$$\text{Percentile rank of a no} = \frac{\text{No of values below than no}}{\text{Total nos } (n)} \times 100$$

$$\frac{2}{10} \times 100 = 20\%$$

20% of the nos are below 3.

* What value exist at 75th percentile?

$$\text{Value} = \frac{\text{percentile}}{100} \times (n+1)$$

$$= \frac{75}{100} \times (10+1)$$

$$= \frac{3}{4} \times 11 = 8.25^{th} \text{ no.}$$

$\frac{75}{100} \times 10 = \frac{15}{2} = 7.5$

$7^{th} + 8^{th} = 7.5$

8.5^{th} no. $= 8^{th}$ no. $+ 9^{th}$ no.

$\Rightarrow 75^{th}$ percentile $= 7$

* mode - highest frequency

1 ~ male

Col.	Col.
1	Male
1	Male
1	Male
0	Female
0	Female
1	Male

→ male

outlier → Cont. numerical data

rank

Good — 1
better — 2
best — 3

1 1 1 2 2 1 2 3 1 1 2 2 5

S. Rs

UP – 20 Rs
↓
2

Nagaland
50 Lakh
↓
2
=