Welcome BacK

Data ⟹ Base of analytics

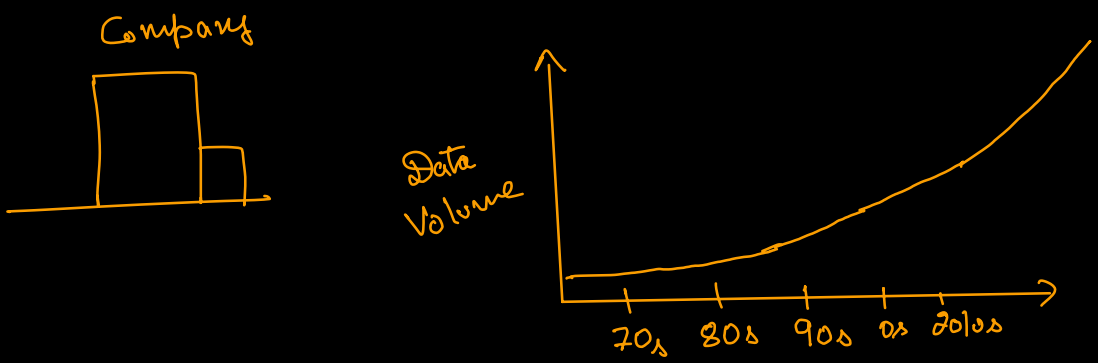
collection of information
any observation

data science

↳ maths.

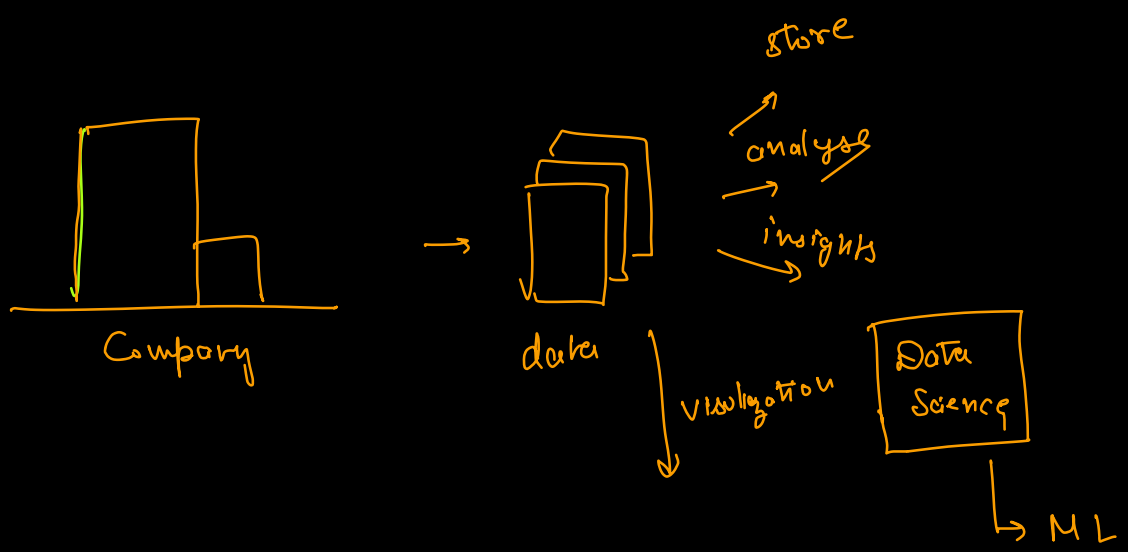
1. Statistics
2. Probability
3. Calculus.

↳ it is a concept / technique used to analyse & derive insights from your data.



If data is getting generated,
we want to use it.

"Data is the new oil"



Tool we can use to analyze / work on data ⇒ Python

It makes money & good for business

1. Statistics
   - Backbone of data science
2. Predictive Analytics
3. Advance Predictive Analysis.

Statistics

Statistics is the science of conducting studies

to collect
organize
summarize
analyze
draw conclusion out of your data.

Statistics deal with collective information data, interpreting this data & drawing conclusion from data

Used in many disciplines: Healthcare
Business
E-commerce
Education
Insurance
Marketing.

Two areas of Statistics

1. Descriptive
2. Inferential

⇒ Descriptive

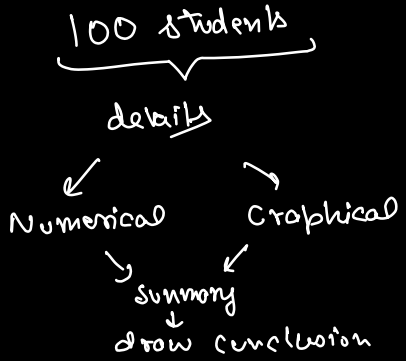
1. help to understand raw data and drive the data acc. to our business problem.

2. Numerical & Graphical to look for pattern in dataset and summarise the data, to present information in a convenient way

Numerical ⇒ mean, median, s.d., variance

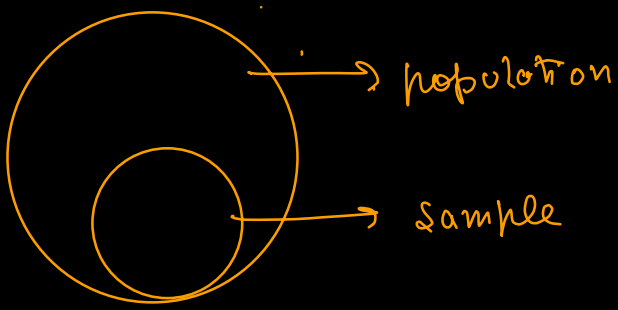
Graphical ⇒ bar charts, histogram, pie charts

GNI PER CAPITA ANALYSIS USING WORLD BANK 2011 DATA	
Mean	11616.7
Standard Error	1361.2
Median	4200
Mode	430
Standard Deviation	17591.1
	3094480
Sample Variance	35.5
Kurtosis	4.99
Skewness	2.27
Range	88680
Minimum	190
Maximum	88870
Sum	1939990
Count	167



Inferential Statistics Inference

⇒ Using sample data to make estimates, decision, prediction or other generalization about a larger set of data.



exit polls

population → 100 people

sample → 20 people

& drow conclusion

POLL OF EXIT POLLS

	SEATS	MAJORITY	PROJECTED RESULTS		
GUJARAT	182	92	BJP 131	Cong + NCP 41	AAP 7
HIMACHAL PRADESH	68	35	BJP 36	Cong 29	AAP 0
DELHI MCD	250	126	AAP 155	BJP 84	Cong 7

Source: Regional and National Channels

Health Warning: Aggregate of exit polls, these polls are not always accurate. Delhi MCD wards 'matched' to ensure a degree of comparability between 2017& 2022 elections

#PollOfExitPolls NDTV

2. With Inferential statistics we are trying to reach conclusion beyond present data

Sample is data ⟶ population's conclusion

Class mean age example

{ 18, 24, 21, 22, 25, 28, 19, 23, 25, 29, 31, 26 }

take 5 people out

18, 24, 21, 22, 25

sample ⟶ mean age of these 5 people

If I say based on above, mean age of full class is something

Exit polls

1. Descriptive

2000 people participated

20 was age mean ⟹ mean of india age is 20

80% had smartphone

Real life phone example

Area B

Apple 35%

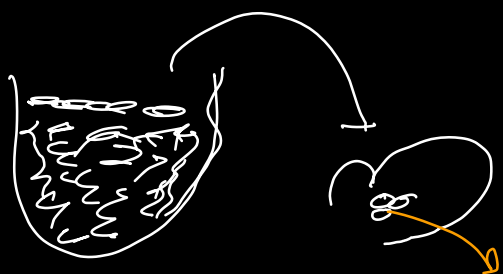
android 60%

other 5%

because they are the same society blocks, with same no. of people with similar char.

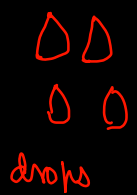
⇒ The mobile distribution will be same

Grain example ⇒



layer for whole population

Blood Test ⇒ of disease.



drops

test them ⟹ infer about full body

Complete data ⟹ descriptive statistics

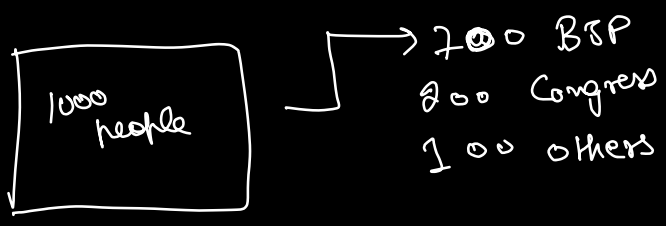
Sample data ⟹ inferring about the population

inferential statistics

don't see data but the end result / use case.

→ only this data — descriptive

→ based on finding statistical quantity, you are inferring for larger population — Inferential



700/1000 people in this 1000 people vote for BJP

these 700 people are rich

70% people will vote for BJP in India

All votes of BJP will be rich

Statistics Jargon

Population : A collection or set of individuals or objects or events whose prop. are to be analyzed.

Infinite or finite.

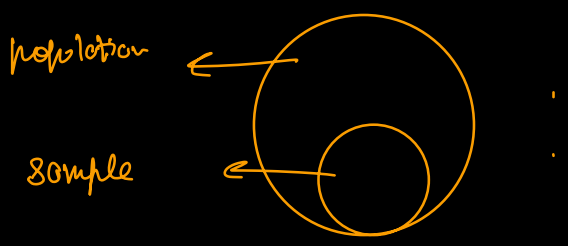
Orders on amazon: 4:00 ——

4:30 ——

infinite pop^{n} when new data keeps on getting generated.

iris
titanic

+ orders on Amazon

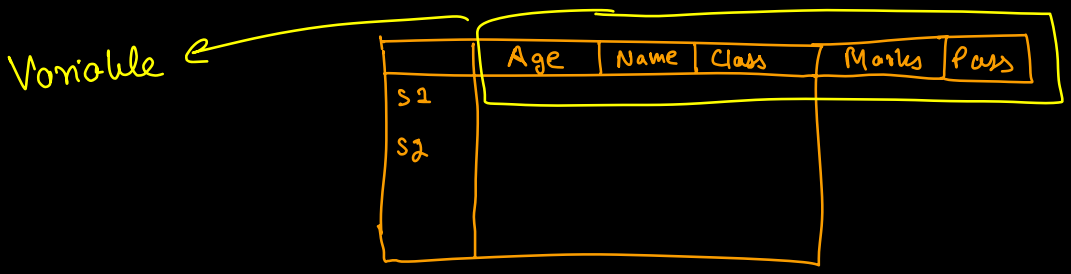


Sample :— A subset of population.

It should be representative of population.

Variable :- A characteristic about each individual element of a population or sample.

Column.



features, predictors, independent variables.

Experiment : A planned activity whose results yield a set of data.

Parameter: A numerical value summarizing all the data of an entire population

Statistic :- Numerical value summarizing the sample data

Variable ⇒ Types of variable

Quantative

measured numericals

no. of runs

marks

no. of students in class

Age of an individual

Height of an individual

population of country.

Discrete

they can take specific, distinct values

eg:

no. of students

no. of classes

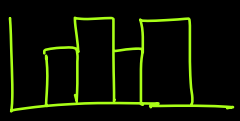
Continous

They can take any values

eg:

Height of students

Temperature



Qualitative [categorical]

Gender → M, F, Other

century

pass or not

eye/hair color

Marital status

level of education.

Dietary preference

Nominal

no intrinsic ordering, purely categorical

⇒ Gender: M

⇒ hair color: B, R, Brown

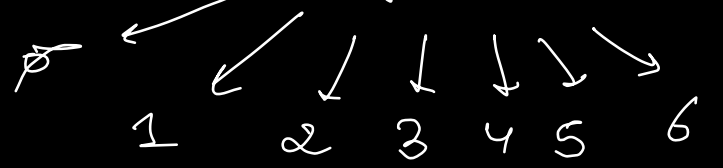
Ordinal

ordering b/w categories

⇒ education qualification

⇒ economic status

O/P of a dice is Categorical



Models of iphone

age

30+

Categorical <= 20 25 30 35 40

age in diapers

age for vaccination

18, 20, 21 ... 80

Can Quantative be converted to Qualitative?

IQ	0 – 50	50 – 100	100+
Smol	not smart	smart	smartes

Money	<10L	10L – 50L	50L+
Economic	Poor	Middle	Rich

Mogenta Red Pink

Pink

Runs by cricketer 1000 players.

runs	0	15	105
⇒	⇓	⇓	⇓

Central Tendency

A measure of central tendency is descriptive statistic that describes the average or typical value for set of observation

- mean
- median
- mode

- std
- variance