Statistics used in Data Science

Definition → Branch of mathematics for collection, analyzing, interpretatin of data & presentation of data.

Importance -

Essential for extracting insights, making predictions and uncovering patterns in data science.

Types of statistics

Descriptive Statistics

↓

summarize the data to provide insight into the data

Inferential Statistics

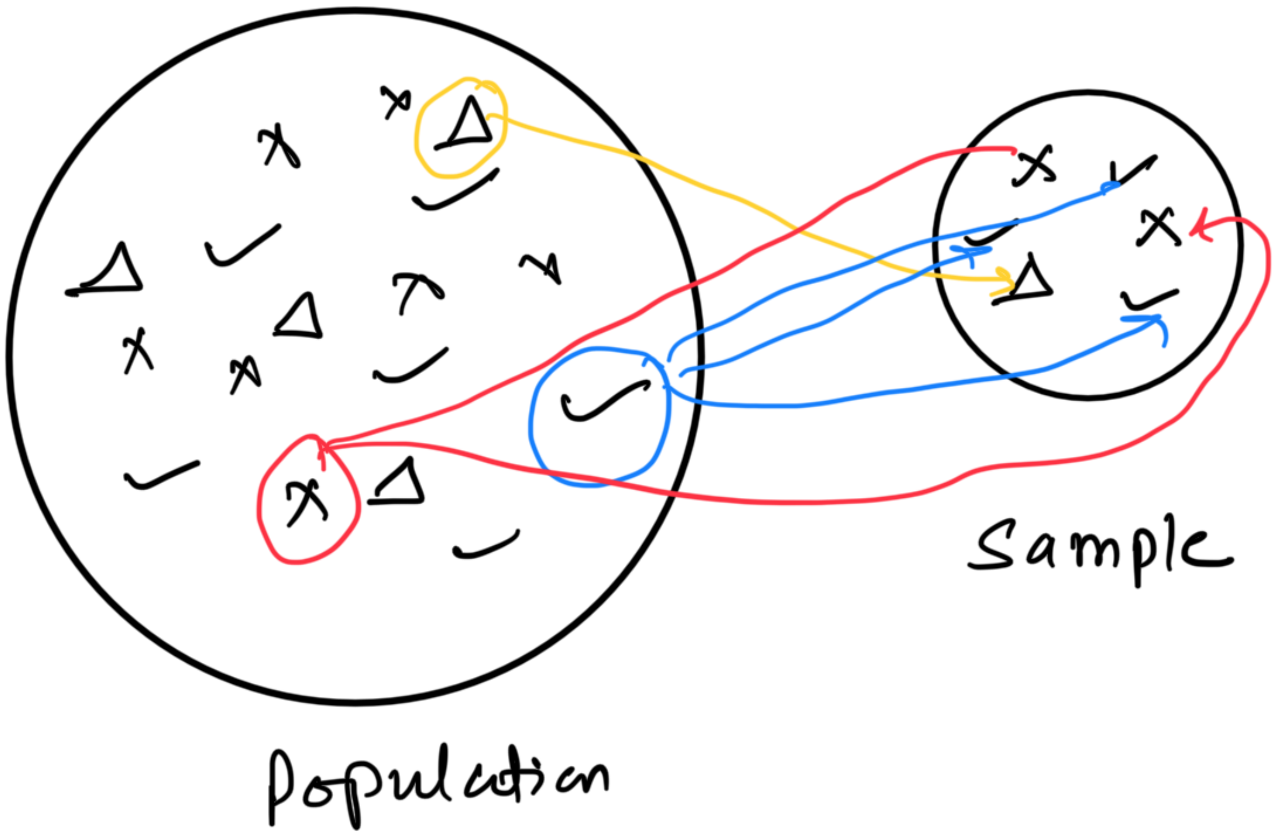
↓

make predictions or inferences about the population based data

Population (N) → is the entire data that you want to draw conclusions

Sample (n) → is the group

(part of population) from which you will collect data



Sampling → Types of Sampling

1. Simple Random sampling -

In this process of Sampling where every member of the population has equal chance of being selected

2. Stratified Sampling -

In this method of Sampling where population (N) is split into non-overlapping group.

3. **Systematic Sampling -**

In this method of Sampling a probability sampling method where researchers select members from population at nth interval

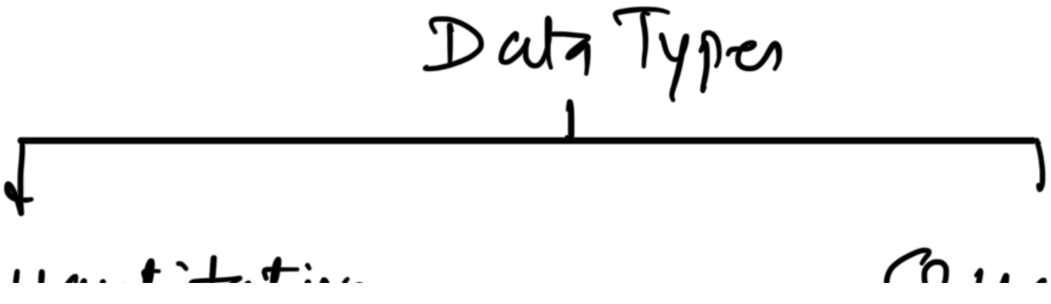
4. **Convenience Sampling -**

In this method of Sampling a process of taking sample data from those who has knowledge / expertise on the domain area.

Data Used in Data Science

Data: Data is a collection of facts, such as numbers, words, measurements, observations or even descriptions of things.

- In the context of data science, data is used to make decision, prediction and inference



Quantitative
(number)

- Take on numeric values

Eg
- No. of students in a class
- Square feet in a house
- Population of a city
- Height of Students

(Measure it numerically)

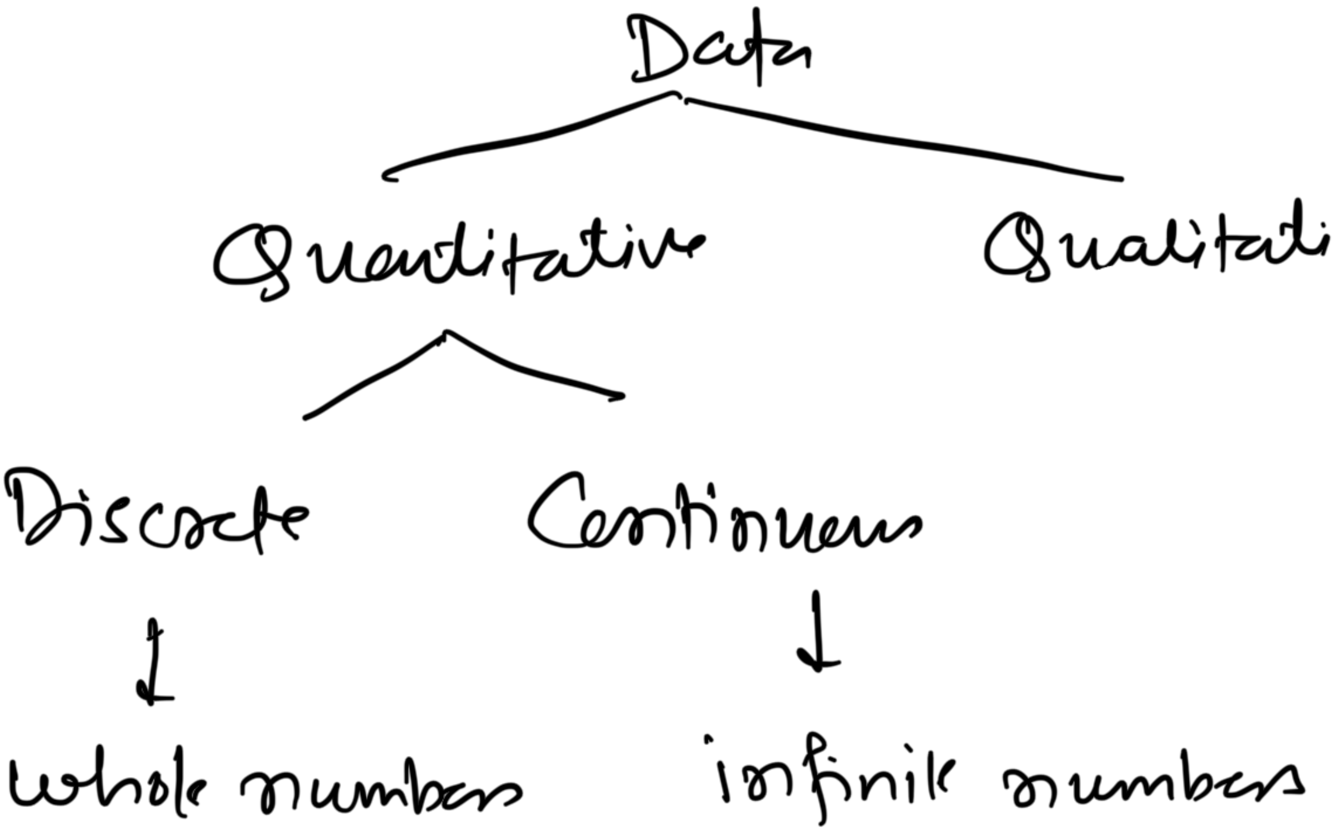
- Add, Multiply, Sub, div

Qualitative
(Categorical)
(textual data)

- Take names or labels

Eg
- Eye color
- Gender
- Breed of dog
- Level of Educatⁿ
- Marital status

(Based on some categorical value)



Discrete Data — Countable and can be taken on some specific value.

eg. No of students in class

No of cars park in parkings

<u>Continuous Data</u> - Can take any value within a range and its often measured rather than count.

eg. Height of students

Temperature Readings

<u>Qualitative Data</u> → Types of data are

1. Nominal ⟶ Categorical data
2. Ordinal ⟶ Order of data matters
3. Interval ⟶ Order matter, {0, max)
4. Ratio ⟶ something measured on ratio scale.

<u>eg. Students Marks and Rank</u>

	St Marks	Ranks	
Pass	100	(first) 1	} → Ordinal data
Pass	96	(second) 2	

Pass	57	4
Pass	85	3
Pass	44	5
Fail	← 30	

Nominal

eg. Temperature

Farehheit

70-80
80-90
90-100

} Interval

eg. Ratio -

Percentile

Data Classifications

Structured Data

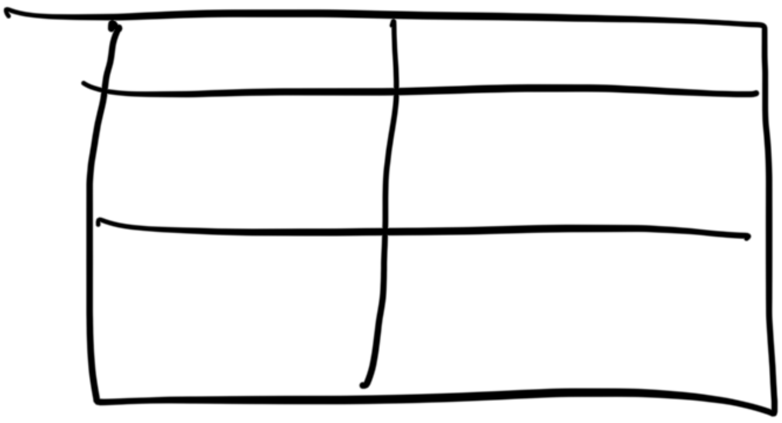
Unstructured Data

Structured Data:

- is organized in a fixed format, often in rows and columns, making it easy for searching & analyzing.
- Eg. Database, Relational Databases

Excel sheets (Spreadsheets)

eg.
```
data = {
  'Name': ['Amit', 'sai', 'Aditya'],
  'salary': ['20k', '30k, '50k']
```



Rows & col
Structured data

Unstructured Data

- data lacks a predefined format, and is not easily searchable.
- It includes huge variety of data formats as text, image, videos, audios, and social meadia post.

eg - Text documents
- Audio files
- Blog post

eg. text data = [" I am kisan."
" Data science is f scinating".
" Python is most popular language."]

Semi-structured data

- It is not organized in a rigid structure like structured data, but it contains tags or markers to separate elements,

- XML language
JSON files

Summary

Data Type	Description	Example
Discrete	Countable specific values	No of students
Continuous	Any value within range	Temperature

Nominal	Categories	Eye color
Ordinal	Ordered categories	Age = { teenage, middleage, oldage }
Interval	Range.	[0 , 100]
Ratio	selected values out of total values	Procentile
Structured data	Organiz in row & Cols	Excel SQL db
Unstructured data	No predefined format	Text docement
Semi-structured data	Tags or markers	JSON files XML

Files