Introduction to Statistics

Statistics / Application

- Descriptive
  - → Central tendency
  - → Dispersion
  - → Symmetricity
- Inferential
  - Probability distribution
    - → pmf
    - → pdf
    - → cdf
  - × CTL
  - Statistical test

Definition : Statistics is a mathematical science including methods of collecting, Organizing and analysing data in such a way that meaningful conclusions can be drawn from them.

Why statistics in data science?

DS/ML → We try to learn pattern in data

* **data :-** facts or pieces of information that can be stored, measured and re-accessed.

* data is used to bring insights to increase a company's revenue.

* Collecting (source of data), Organising (reaccess the data conveniently) & analysing.
  - Collecting
    - ↳ Portal
    - ↳ manually
    - ↳ surveys
    - ↳ Internet clicks
    - ↳ Experiments
    - ↳ directly from customers through rating & reviews.
  - Organising
    - ↳ database
      - eg: SQL
      - NoSQL.
  - analysing
    - ↳ finding insights from data using tools like pandas, Numpy, matplotlib, seaborn and so on.

Examples → scores made by Virat Kohli in T-20 world cup

[0, 1, 9, 10, ... 74]

→ Bumrah Economy rate

→ Rohit Sharma in t-20.

→ Demand of a product for an ecommerce site.

→ Height of students in the class.

Motivation / Uses of Statistics

1. Weather forecast.
2. Sports Analytics ⟶ HS, SR, MIV.
3. Election Campaign
4. FMCG | E-commerce.
5. medical / genetics / pharmaceutical.

Types of Statistics

① Descriptive Statistics

⇓

Describe

Defⁿ → It consists of Organizing and summarising the complete data / population.

Q: How good Rohit Sharma is as a cricketer?

R.S → All the match scores of his entire career and then calculate Avg score.

[4, 5, 0, ... 264 ...]

② Inferential Statistics

↓

Inference

→ You can not count no of trees of a jungle

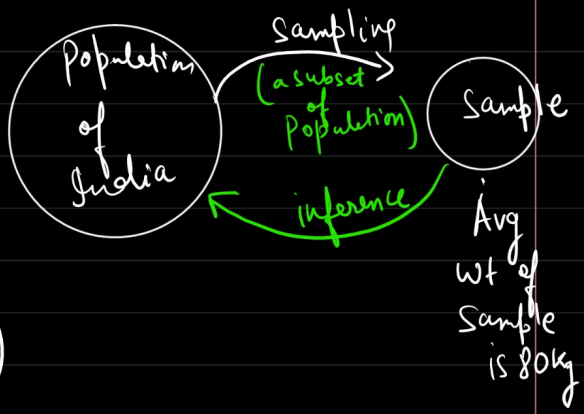
→ You can not go to 150 Cr+ People of India to record their weights

⇓

Inferential Stats

Defⁿ It consists of using sample data that has been measured to form conclusion about a population.

→ with given sample data can we conclude something about population



① Descriptive Statistics (population summarisation)

Example 1 → SR of Virat Kohli

Example 2 → Avg ht/wt of students in this class | state | religion | country

Example 3 → CEO of an airline company ⇒ Avg delay of flights of your company

* Whenever you want exact statistics to take any business decision ⇒ descriptive statistics

Q: What is avg height of students in class? (150+160+170...)/avg

Q What is the most frequent heights of students in class? (170cm)

Q What is average delay of flight of abc airline?

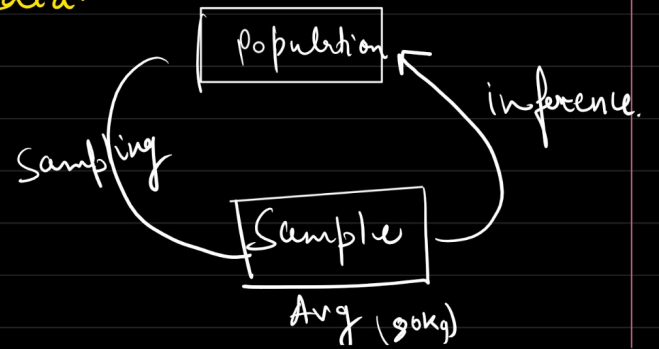
Techique of Descriptive statistics

1. Measures of Central tendency (mean, median, mode)
2. Measures of Dispersion (Standard dev^n, Variance)
3. Measures of Symmetricity (Skewness, Kurtosis)

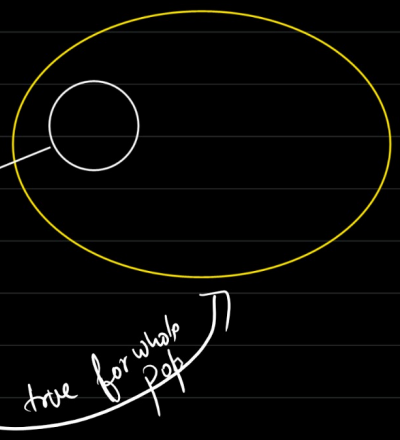
② Inferential statistics (conclusion for population made from sample)

Why? → Population is large
→ time & resource constraints.

Eg: Avg age | weight of population of India.



eg. No of trees in a National park.



10% of area = 2000
100% of National park = 20,000

Sample area → Count

Summary

Descriptive

Classroom
→ what is Avg height / weight of students in class?

Inferential

→ if the avg height of students in a specific class (say class 6 as sample), What you expect for the entire school.

150cm

* Taking Sample is important in Inferential statistics.

⇓

Sampling techniques.