Lifecycle of a DATA Scieme Projects.:

① Requirement Gathering → ② DATA ANALYTICS TEAM: Data Analyst, Data scientist → will find how to gather data to solve the problem

BUSINESS ANALYST ↑ ↑

PRODUCT MANAGER ⇅ PROJECT MANAGER (Domain Knowledge) } discuss with product manager to know the requirement

① Data Analyst
② Data Scientist
③ BIG DATA Engineer
④ Cloud Engineer

↓

③ Sources of Data

Internal Database | 3rd Party API's | Web Scraping

↓

Big Data Engineering → save to → MySQL or NoSQL → Database

④ Data Scieme Projects → Data Scientist works starts here

① EDA → Need statistics
→ Analysis of data done by drawing plot.
scatter plot

descriptive Stats ⇒ Summarizing the Data
Pie chart


↓

② Feature Engineering → Need Statistics

↓

③ Feature Selection → Need Statistics

↓

④ Model Training → Trained with ML algorithms.
⇒ Need statistics

↓

⑤ Hyperduning → Improve the performance of the model

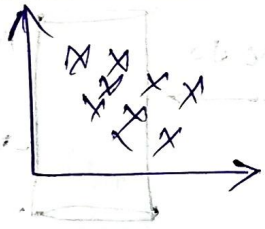
↓

⑥ Deployment

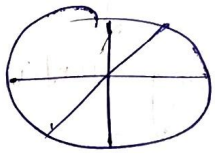
Analysis of Data

① Age = { 12, 13, 14, 15, 20, 25 }
Find average age?
→ This is also descriptive statistics.
Measure of centered tendency.

② Scatter Plot



Pie chart



← Descriptive Statistics →

Statistics

→ It is the science of collecting, organizing and analyzing of data.

Data

→ "Facts" or pieces of information".

Eg: ① Ages of students in classroom
{ 24, 25, 22, 29, 28 } ⇒ Mean, Median, Mode, standard deviation

② Weights of students in classroom
{ 55, 65, 72, 73, 89 } ⇒ Mean, Median, Mode, standard deviation

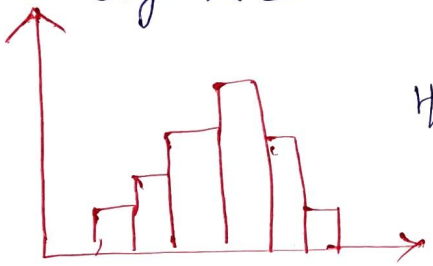
Types of Statistics

1. Descriptive Statistics
2. Inferential Statistics

Descriptive statistics

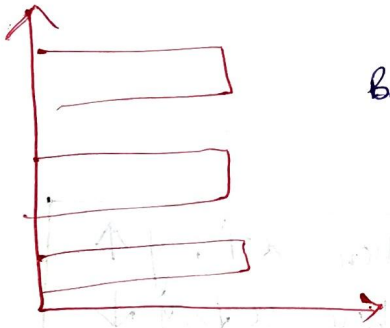
→ It consists of organizing and summarizing of data using different plots.
→ Extensively used in EDA and Feature engineering.

Eg: ①



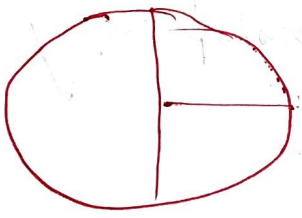
Histogram

②



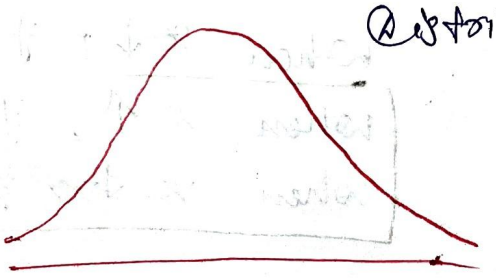
Bar chart

③



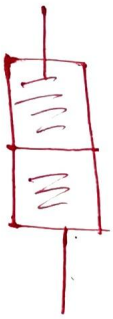
Pie chart

④



Distribution

(5)



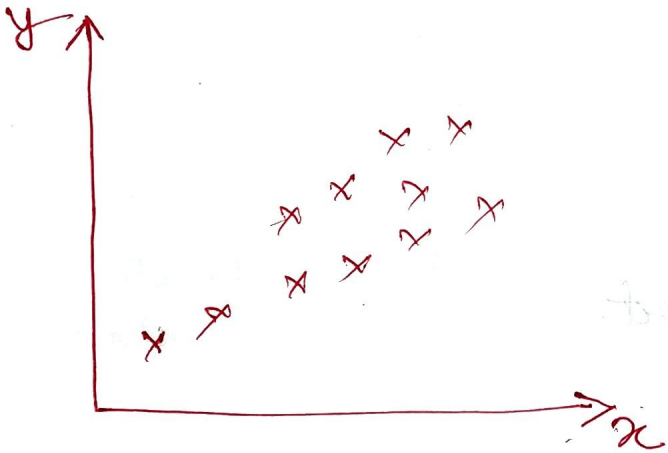
Candlestick

(6)

Box Plot

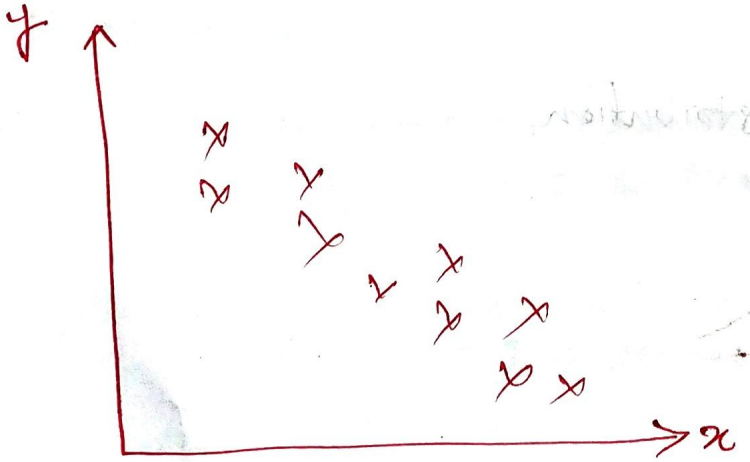
(7)

x (Height)	y (Weight)
—	—
—	—
—	—
—	—
—	—



when $x \uparrow, y \uparrow$
when $x \downarrow, y \downarrow$.

→ We can check this relationship using scatter plot.



~~when $x \downarrow, y \downarrow$~~

when $x \uparrow, y \downarrow$
when $x \downarrow, y \uparrow$

Inferential Statistics

→ It consists of collecting sample data and making conclusion about population data using some experiments.

→ Making conclusion → can be done by hypothesis testing.

Eg: ① University → 500 students

class A → 60 students

⇓

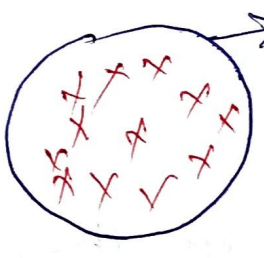
Sample data → Age → Average age of the entired universities.

Can we make average age of the university with the help of sample data of class A?

→ Yes, we can make the prediction.

Sample Data → Conclusion about → Population Data through Hypothesis testing.

Sample Data Vs Population Data



→ Punjab

Total Population = 10 cr

Exit Poll.

→ They have to take sample of data.

Sample size = 1000

Take sample size from every region and then do average and comes up with exit poll.

⇒ PARTY A will win (Earlier/Prediction)
PARTY B will loose. (Actual result)

⇒ Prediction goes wrong, hypothesis testing goes wrong.

Problem:

Let's say there are 20 classrooms in a university and you have collected the age of students in one classroom?

Ages = {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22}

Weights = { - - - - - - - - - - - }

⇒ Descriptive Stats:

Ans: ① What is the average age of students in the classroom?

② Relationship b/w age and weight?

Inferential Stats:

Ans: ① Are the average of the students in the classroom less than the average age of the students in the university?
comparing sample and Population data.

②

② 1000 → students

class A → 50 girls → 95% marks

class B → 50 boys → 92% marks

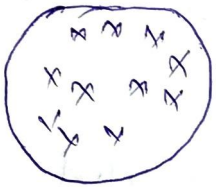
Are the average marks of by girls in the university greater than by boys?

⇒ We can do this using Hypothesis testing.

Sampling Techniques

① Simple Random sampling

⇒ Every member of the population (N) has an equal chance of being selected for your sample (n).

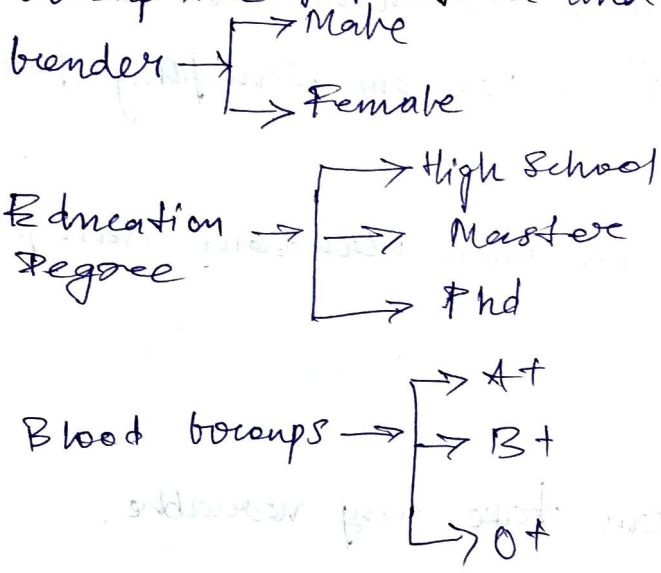


⇒ We can perform random sampling for exit poll, General Survey etc.

② Stratified Sampling

Strata → Means layers → clusters, → groups.

⇒ Group the population and then select.



E.g: Exit Poll ⇒ age ≥ 18 and age < 18 (Can't vote)

③ Systematic Sampling

{AIRPORT}

CREDIT CARD sale place

⇒ Select every 9th individual out of population.

(4) Convenience Sampling

→ only those who are interested in the survey, will only participate.

E.g: "x" people is interested in Data Science survey. We will send survey documents to x people only.

(2) iNeuron job form → fill the form (Interested candidates)

Que.

(1) Survey regarding new technology?
→ Convenience Sampling.

(2) RBI survey for women to know the expenses?
→ Stratified Sampling, then Random Sampling for married women.

(3) Credit card calls?
→ Stratified Sampling and then random sampling. (for salaried person)

Variable

→ It is a property that can take any variable.

eg: age = 14
age = 33
age = 100

Variables

Ages = [24, 25, 26, 27, 28, 29] ⇒ collection

Types of Variable:

1. Quantitative Variables
   - Discrete variable
   - continuous variable
2. Qualitative Variables
   - categorical variables

Quantitative Variable

→ Measured numerically & Mathematical operations §

e.g: Age, weight, height, rainfall (cm), temp, distance.

Qualitative Variable

→ categorical variables § Based on some characteristics they are grouped together §.

e.g: Gender, Types of flowers, Type of movies

Quantitative Variable

① Discrete Variable

→ whole number.

e.g: ① Number of Bank account
② Number of children in a family

② continuous variable

→ continuous, decimal

e.g: ① Height
② Weight
③ Rainfall
④ ~~Up~~ Speed

Assessment

① What kind of variable is Marital status?
→ Categorical variable

② What kind of variable is Orange River Length?
→ Continuous

③ What kind of variable is Movie duration?
→ Continuous

④ What kind of variable is Periode?
→ Discrete (Periode has many varieties). that's why it is not categorical.

⑤ What kind of variable is IQ?
→ ~~Continuous~~ Discrete (IQ doesn't has 'no' names)

⑥ What kind of variable is Gender?
→ Can Categorical (Gender has few limited categories)

[→ Note: Many number of categories → Put it into Discrete]

⑦ What kind of variable is Pancard number?
→ Categorical (Pancard number has set of character)

Pancard has alphanumeric character → Categorical feature and it is also a unique identifier