DATA SCIENCE

# Statistics for Data Science

03/09/2022

## 1. Usecase:

Bank → Let Suppose HDFC

(A) (C) (B)

This task is Basically for Data Analyst or Data Scientist.

Q HDFC have already two ATM at Location (A) & (B) They are Planning whether they will open new ATM at Location (C) or not?

## 2. Use case :-

Q. Find the Average Size of Shark throughout the world?

## 3. use case :-

Q. Amazon Big Billion Day Sale
(this Question asked in Intuit (Product Based) company)
* Which Month should you select for the Big Billion Day Sale?

## Statistics [Life cycle of Data Science Project]

# Life cycle of a Data Science Project :→

① Requirment Gathering → ② DATA ANALYTICS TEAM: DATA ANALYTICS or DATA SCIENTIST → will find how to gather data to some problem

Business Analyst

② Product Manager or Project Manager (Domain Knowledge or Domain EXPERT) — Discuss with Product Manager to know the Requirment

① DATA ANALYST
② Data ScienTist
③ Big Data engineer
④ Cloud engineer

③ Sources of Data

- Internal Database
- 3rd Part APIs
- Web Scrapping

→ Big Data Engineering

→ DATA BASE → MySql or NO Sql

My-SQL

⇒ We use to save text in My SQL.

NO SQL

⇒ For saving the images we use NO SQL

(4) DATA SCIENCE PROJECTS:-

⇒ The work of DATA Scientist starts here.

(i) EDA :-> Exploratory Data Analysis (Need Statistics)

(ii) Feature engineering

(iii) Feature selection

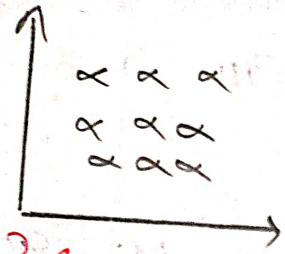
(iv) Model selection -> Trained with ML Algo

(v) Hyper Parameter tuning -> improve the performance of Model

(vi) Deployment

* In All the steps Statistics will be used !!

# Analysis of Data



Summarisation of Data

↓

Descriptive Statistics

Pie chart

Descriptive Stats.

Age = {12, 13, 14, 18, 20, 25}

Average age {is Part of Descriptive Stats} is a Measure of central tendency

Interview Question

Q. How Statistics is used in Machine Learning?

## Definations

### Statistics

=> Statistics is science of collecting & organising & analysing the Data.

### Data

=> Data is nothing but fact or piece of information

Eg:-

(1) Ages of students in classroom {24, 25, 32, 29, 18}
with the help of this Data we can analyse this Data by finding Mean, Median, Mode & Standard deviation

(eg 2) weight of students in a classroom

---

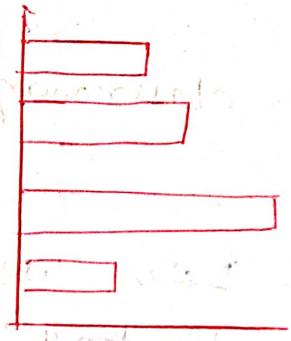
## Types of Statistics

- Descriptive stats
  - -> It consist of organising & summarizing of Data
  - -> Extensively used in EDA & feature engineering
- Inferential stats
  - => It consists of collecting sample Data & making Conclusion about Population

# Descriptive Statistics

1. 

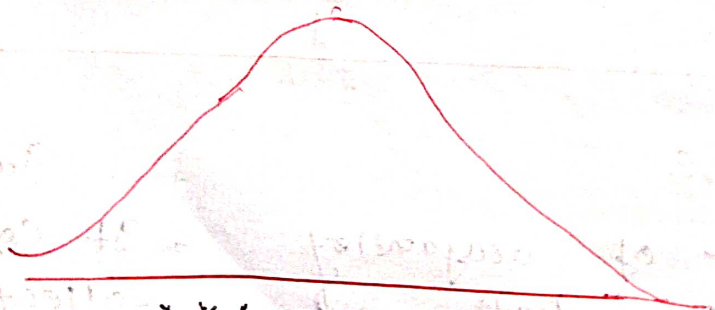
Histogram

2. 

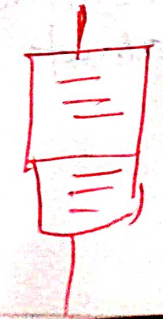
BAR CHART

3. 

Pie Chart

4. 

Distribution

5. 

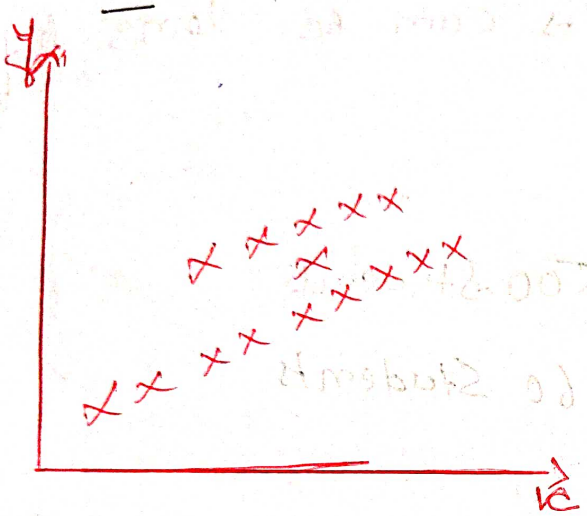
Candle Stick

6. 

Box Plot

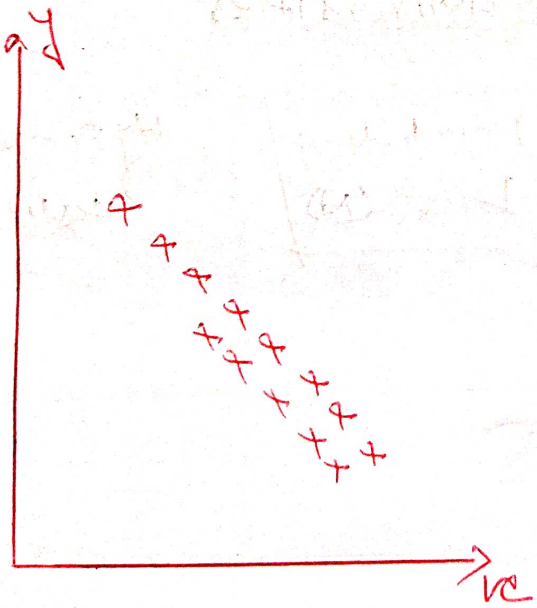
(7)

| x | y |
|---|---|
| Height | weight |
| — | — |
| — | — |
| — | — |
| — | — |
| — | — |
| — | — |



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

→ we can check this Relationship using Scatter Plot



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

# Inferential Statistics

→ It consist of collecting Sample Data & Making Conclusion about Population data using some experiment

→ Making conclusion → can be done by hypothesis testing

Eg: (1) University 500 students

Class A → 60 students

↓↓

Sample Data ⇒ Age ⇒ Average age of entire universities

| Sample Data (n) | CI Confidence interval | Population Data (N) | Hypothesis testing |
|---|---|---|---|

P-value

(1) Z-test

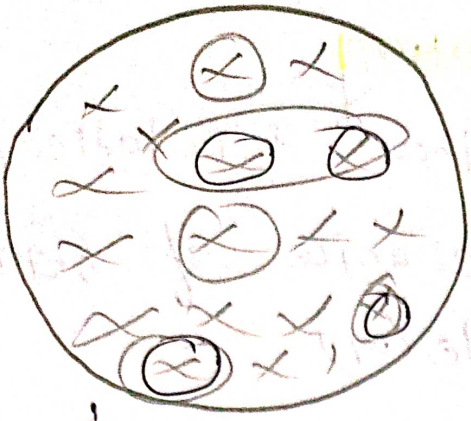
(2) t-test

(3) Chi-Square test

(4) F-test

# Sample Data vs Population Data

(n) (N)



Punjab State
(10 Cr. Population)
↓
Population Data

Newchs will conduct some exit poll

We can't go to every people to ask, so we will select some sample of size 1000 & ask from that peoples

Eg:→ Let's say there are 20 classrooms in university & you have collected the age of students in one classroom.

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

Weight = ?

}

Descriptive Stats

Q. What is Average age of student in classroom?

& Relationship Between Age & weight

Inferential Stats

Q. are the average age of the student in the classroom less than Average age of the student in the university?

Population Data = (N)

Sample Data = (n)

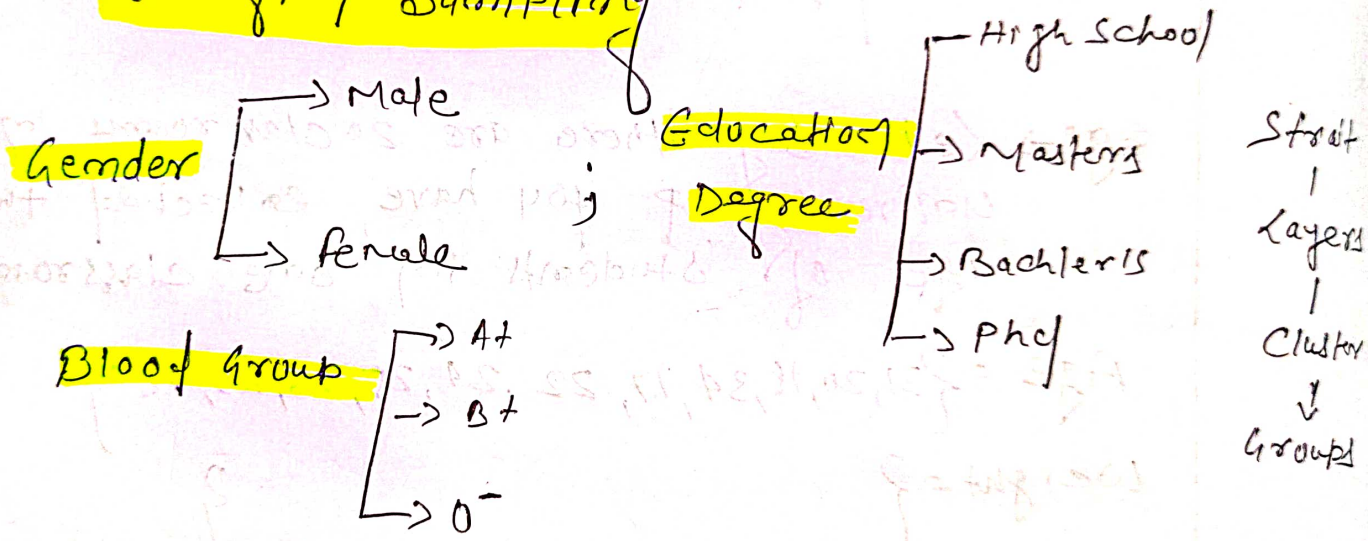
Q. What are the different Sampling technique?

1. Simple Random Sampling

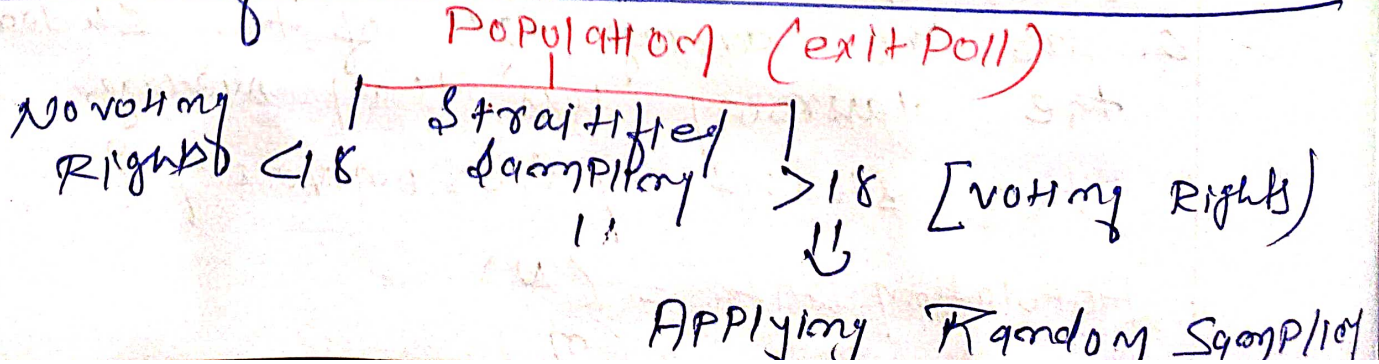
⇒ Every member of the Population (N) has equal no. of chance of being selected for Sample (n)

Eg:→ exit Polls, Movie Review, Lottery

2. Stratified Sampling



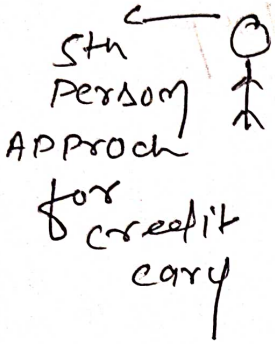

Eg exit Poll, we first Apply Stratified Sampling for neglecting age <18 & choosing the age >18

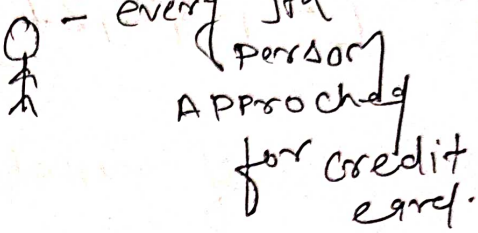


## ③ Systematic Sampling

⇒ Selection of every nth individual out of Population (N).

Example :⇒ { At MALL }




## ④ Convenience Sampling

⇒ Only those who are interested in survey. only they will Participate

Example :-

① Survey Regarding New technology

② Credit card calls.

③ RBI Survey for know the House expense only for women — Married / Stratified / Unmarried → convenience Sampling

## Variable

⇒ A variable is Property that can take any values

eg:- age = 14, 25, 100

## Variables

Ages = [24, 25, 26, 27] ⇒ Collection

## Two Types of variable

(1) Quantitative Variable

:⇒ Measured Numerically (Mathematical operations of)

eg:- Age, weight, Rainfall (cm), temperature, Distance

(2) Qualitative variables

Eg:- Gender, Type of flower, Type of Movie

* Based on some characteristics they are grouped together

Quantitative Variable

↓

Discrete (whole No) variable

* No. of Bank Account
* No. of children
* No. decimal value

Countinous variable

eg:- Countinous

(1) Height, weight, Ages, Rainfall, Speed.