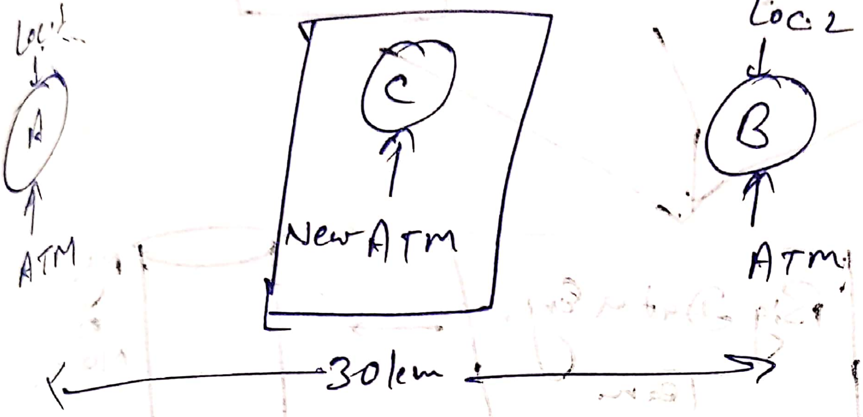
Statistics

18/12/2022

Usecases.

1.) HDFC ← Asked in Product Based Company?

Check whether to open this ATM or not?



Can work as
→ Data Analyst
→ Business Analyst
→ Data scientist
→ Product Manager

This task can be solved by
① Data Analyst
② Data scientist

or Blue whale ← (Amazon)

2.) Find the Avg. size of the shark throughout the world.

(Intuit)

3.) Amazon Big Billion Day sale. {~~Intuit~~} → which month should you select?

Topics to cover in statistics :- {Life cycle of Data Science?}

Data science ↓ ↓ ↓ ↓

Requirement Gathering → Data Analyst / Data Scientist

Data Analytics Team

Product manager ⇑ Project Manager

Business Analyst

① Data Analyst
② Data scientist
③ Big Data Engineers
④ Cloud Engineers

Domain knowledge ⇑ Product manager → Data Analyst

BA (Business Analyst) → Data Scientist

Project

Types of company

Service - TCS, Infosys,

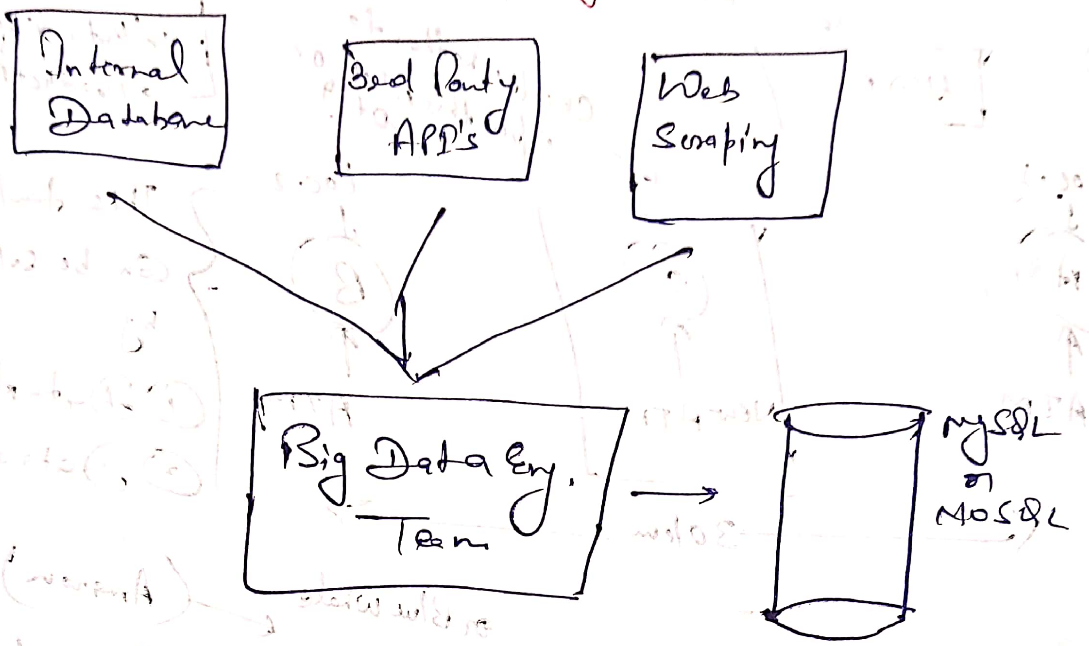
Product Based

(IBM, {Google}, Apple,

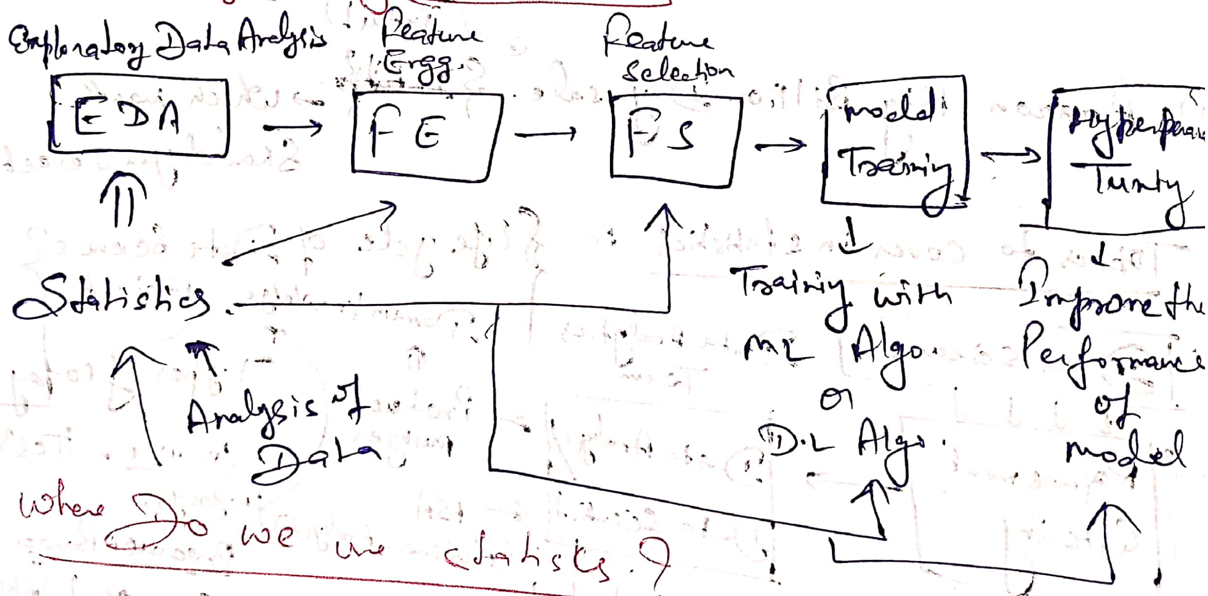
YT, G Pay, Google Ads, Gmail

Sales ↑ Domain experties ⇓ Product manager

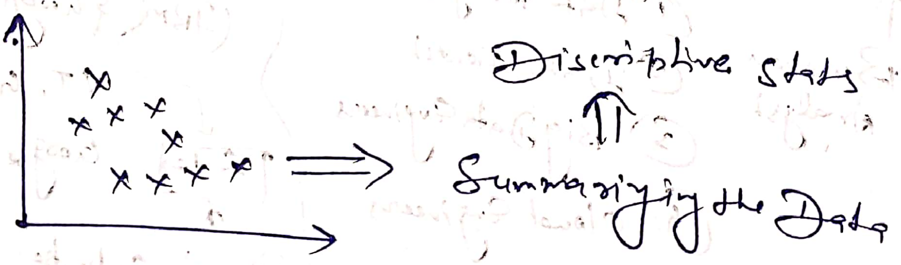
From Where Can You get the Data

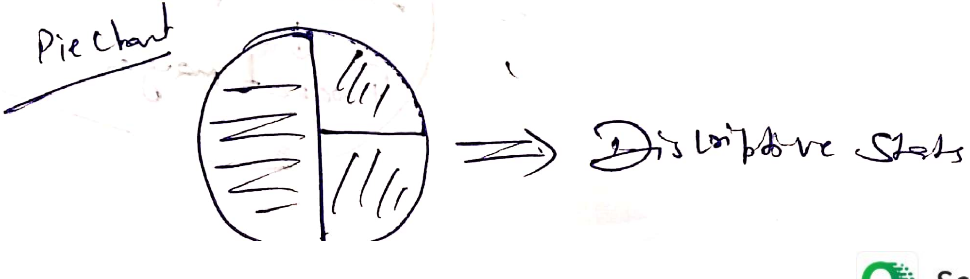


Life Cycle of DS Project



Where Do we use statistics?





Age : {12, 13, 14, 18, 20, 25} ⇒ Avg. Age }

⇓ → Measure of central Tendency

Part of Discriptive stats.

Stastics Def^n :→ Statistics is the science of collecting, organizing and analysing the Data.

Data :→ "facts or pieces of information"

Eg: (i) Ages of Students in classroom.

{24, 25, 32, 29, 28} ⇒ Mean, Median, Mode, Standard Devistion.

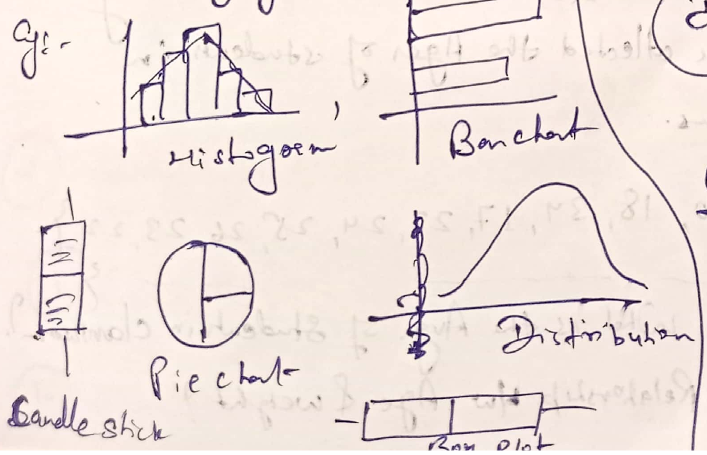
(ii) Weights of Students in classroom.

Type of Statistics

Descriptive stats

(i) It consists of organising & summerizing the data.

Eg:-



Inferential stats

(*) It consists of Collecting sample data & making Conclusions about Population Data using some experiments.

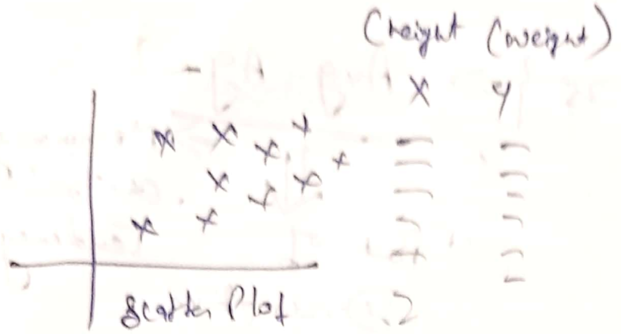
→ Hypothesis Testing.

University

Class A → 60 People

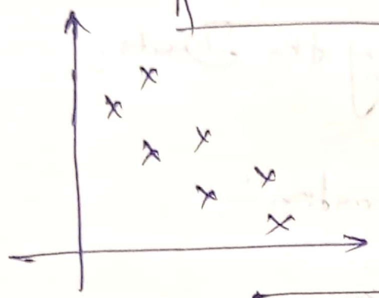
⇓

Sample data ⇒ Age ⇒ Avg Age of entire university



X ↑ Y ↑

X ↓ Y ↓



X ↑ Y ↓

X ↓ Y ↓

Hypothesis Testing

C.I ⇒ Confidence Interval

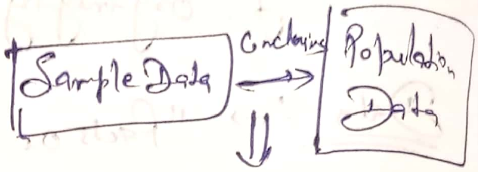
P- value

① Z - Test

② t - test

③ Chi square test

④ F Test (Annova Test)



Hypothesis Testing }

Sample Data Vs Population Data.

10cr ⇓ Population Data

Sample size = 1000

Exit Poll ⇑

Party A will win
Party B will lose } If opp. happens means hypothesis is gone wrong.

Eg:- Q Let's say there are 20 classrooms in a university and you have collected the Ages of students in 1 classrooms.

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

Weight = { ___ }

Descriptive stats :→ What is the Avg. of Students in classroom?
Relationship b/w Age & weight ?

Inferential stats :→ Are the avg. age of the students in the classroom is less than the avg. age of students in the university

⇓ greater than

⇓ equal to

Eg:- • University has 1000 Students

Sample Data

⇓

Class A ⇒ 50 girls ⇓ Avg. marks 95% 50 boys ⇓ Avg. marks 92%

Q Can we conclude that Girl has done better than Boys in the entire university?

⇒ yes for this we have to Apply Hypothesis Testing.

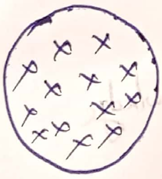
Population denoted by (N)

Sample denoted by (n)

⇒ Different Sampling Techniques :→

① Simple Random Sampling (SRS) :→ Every member of the population (N) has an equal chance of being selected for your sample. we can use Simple Random Sampling

Eg:-



Exit:→ Poll

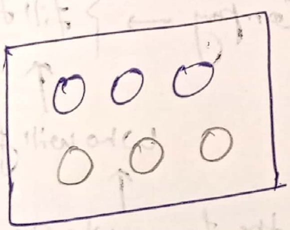
General Survey,

Movie Reviews,

Lottery.

N = 2,

②



n = 3

Every marble has equal chance of getting selected

② Stratified Sampling :→ Strata → Layers → Clusters → Groups

(i) Gender → Male, Female

(ii) Education Degree → High school, masters, PHD

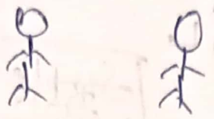
(iii) Blood groups → O+, A+, B+

(iv) Exit Poll → <18, >18 ✓ ← Apply Random sampling.

③ Systematic Sampling :→ {AIRPORT}

{Credit Card}

Approach every 5th Person for credit Card



Approach every 9th Person for ~~Appr~~ credit Card

Here we select every nth individual out of Population (N).

Eg: every 5th Person, or every 10th Person etc.

they are following some systematic ways.

④ Convinience Sampling :→ Only those, who are interested in the survey will only participate

Eg:-

(i) { Data Science Survey :- General AI Survey . }

⇑

Whoever Participate in this survey Should Be interested or have knowledge about Data Science

Eg (ii) iNeuron Job for specific Company → {fill the form}

↑

Who will fill the form?

↑

Those will fill the form who are interested for that Particular job.

Q. What Sampling, Can we use in the following Situation?

(1) Survey regarding New Technology → Convinience Sampling (Bcz only those who are interested technology will Participate)

(2) RBI Survey by → women → Stratified Sampling + Random Sampling

(Since women takes Care of entire house)

women → Married women → Random Sampling

(3) Credit Card Call :→ Stratified + Random (Bcz, mostly they calls Salaryedpeople & then Random)

⇒ Variable :→ It is a property that can take any value.

Eg:- age = 14
age = 25
age = 100

Variables
Ages = [24, 25, 26, 27, 28, 29]
↙ list, collection

⇒ Two Different types of Variables :→

(1) Quantitative Variable :→ Measured Numerically {Mathematical operations}

Eg:- Age, weight, height, rainfall, temp, distance

(2) Qualitative Variables (Categorical Variables)
↓
Based on Some (Bcz of Characterics they are grouped together)

Eg:- Gender, Types of flowers, Types of movies etc.

Quantitative Variables (Continious Variab

Discrete Variable

eg:- Whole no. → fixed (decimal nos. not allowed)

i) No. of Bank a/c.

{1, 2, 3, 4, 5} 2.5 ✗

ii) No. of children

Should have less no. of categories.

Continious Variable

Contins → Decimal Value

eg:- Values will be → Any value.

eg:- Height, Weight, Ages, Rainfall, Speed etc.

Can have Whole nos. also.

⇒ Assessment

① What kind of Variable is Marital status? → Categorical Var
→ married
→ Not married

— — — — Ganga River Length? → Continious

— — — — Movie Duration? → Continious

— — — — Pincode? → Discrete ~~105.75, 90.5~~

— — — — IQ? → ~~Contiriom~~ Discre

Since it doesn't have decimal values.

— — — — Gender? → Categorical

— — — — No. of People married? → Discrete.

Pincode: 160099, 560098, 560097 } It is many? so we can't take it as Categorical.

final

Categorical ⇓ Feature Engg.

Unique PANCARD → Categorical Variables

Since it is having Alphanumeric values.