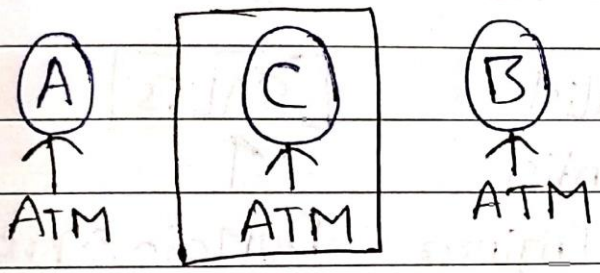
: Hare Krishna :-

Statistics:-

Use case:

HDFC

30 KMS



Statistician

↳ 5 years

- ↳ Data Analyst
- ↳ Bussiness Analyst
- ↳ Data Scientist
- ↳ PRoduct MANAGERS

1. DATA ANALYST
2. DATA SCIENTIST

Product Based company & Service Based company

Product based : {Google} → G.Pay, GMAIL, youtube, chorem

Apple, IBM, Microsoft, intel

Service Based:

TCS, HCL, infosys

MATRIKAS

Date: / /

Statistics:-

Life cycle of data Science

DATA SCIENCE

Projects

Domain Knowledge

Requirement Gathering → Data Analyst / Data Scientist ← Product Manager ← BA

Business Analyst

Product Manager ⇕ Project Manager

1. Data Analist
2. Data scientist
3. Big Data Engineer
4. cloud Engineer

SALES

Domain Expertise

Product Manager

internal Data base

3rd Party API's

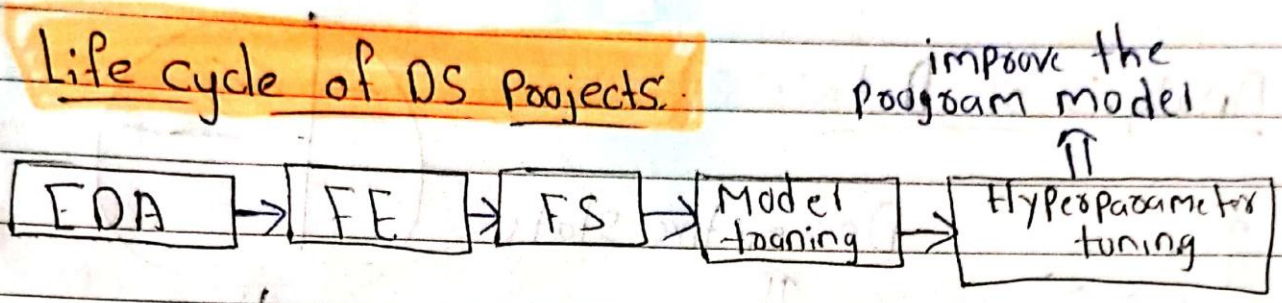
Web scraping

Big Data Eng →

MySql or NoSql

Date / /

Life cycle of DS Projects.



EDA :- Exploratory Data Analysis
FE :- Feature Engineering
FS :- Feature Selection

Statistics are used in

→ EDA
→ FA
→ FS
Training with ML Algorithms ⟺ → Model training
→ HyperParameter tuning.

improve the Project model

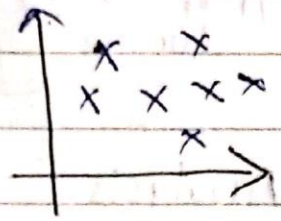
Product base companies:- focuses on making the highest quality product for its customers
Ex:- Microsoft, Google, cisco, Adobe, Apple ... etc

Service based companies:- Provide required service to the another company Products
Ex:
TCS, infosys, IBM, Wipro, Accenture ... etc.

MATRIKAS

Date / /

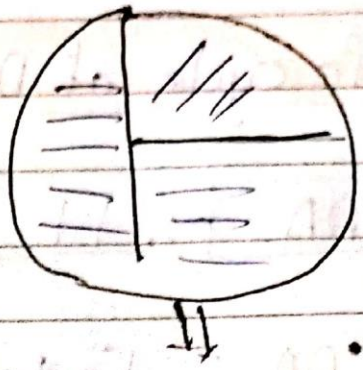
Analysis of Data:



Descriptive stats

⇑

Summarizing the data



Descriptive stats

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

⇓

descriptive stats

Measure the central Tendency

Defination of Statistics:-

Statistics is the scienu of collecting, organising and analysing the data.

Data:- "Parts or Pieces of information"

Eg:- Ages of students in classroom

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

Standard deviation

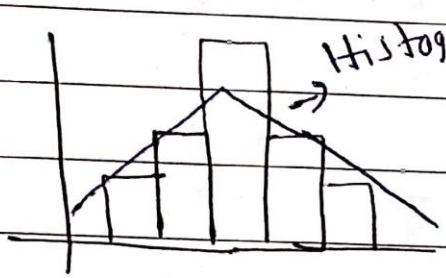
MATRIKAS

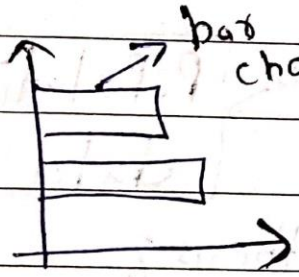
Date: / /

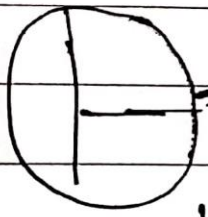
Statistics

Descriptive stats [EDA + FE]

① it consists of organising and summarizing the data







Box Plot.

Inferential Stats

① it consists of collecting sample data and making conclussion about population data using sum Experiment

Hypothesis testing

university → 500 people

ClassA → 60 People

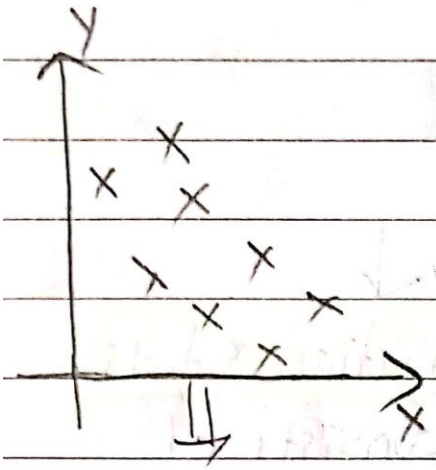
Sample data ⇒ Age

Average age of entire university

Hypothesis testing

C.I ⇒ Confidence interval

Date: / /



(height) (weight)

x y

\- -

\- -

\- -

$x\uparrow\ y\uparrow$

$x\downarrow\ y\downarrow$

Scatter plot

$x\uparrow\ y\downarrow$

$x\downarrow\ y\uparrow$

P-value

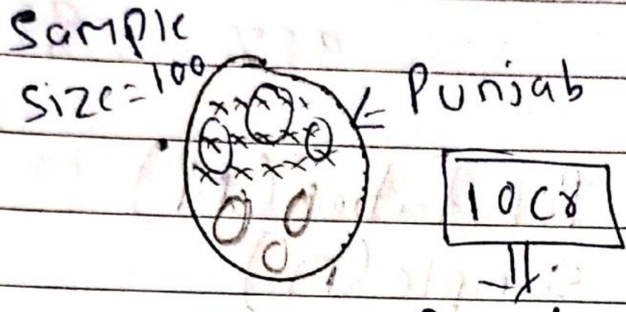
1. Z-test
2. t-test
3. Chisquare test
4. F-test

Sample Data	conclusion ⇓	Population data

Hypothesis testing

Date: / /

Sample data & Population data:



Exit Pole:
Party A will win
Party B will loose

Eg: let say that 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 }
Weight = { - - - - - - - - }

Descriptive stats:

What is the average age of students in the classroom?

Relationship b/w Age & gender?

inferential stats:

Are the average age of the students in the classroom less then the average age of the student in the University

less
not
the equ

Date / /

1000 students ⇒ hypothesis testing

Class A ⇒ 50 girls 50 boys

= =

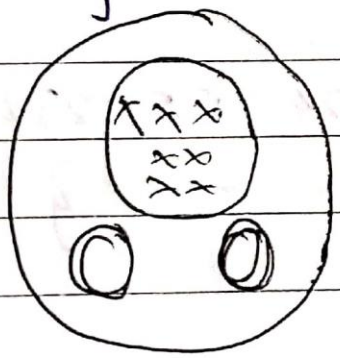
95% 92%

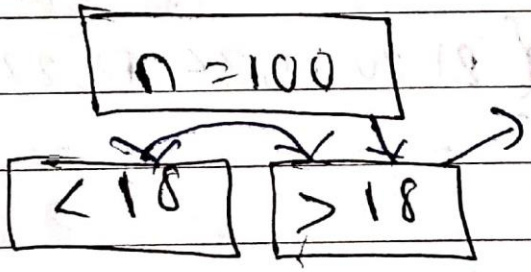
Choose a sample :-

Sampling Techniques : Population (N)
Sample (n)

① Simple Random Sampling :

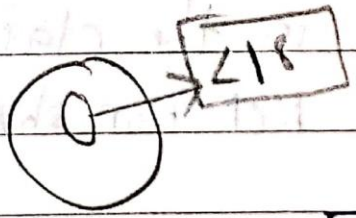
Every member of the Population (N) has an equal chance of being selected for your sample (n)





② Stratified Sampling :

Strata → layers → clusters → groups



Gender → male, female

Education Degree → high school, master, phd

blood group → O +ve, B +ve, B -ve, not vote

Population → < 18, ≥ 18

{Exit Poll}

MATRIKAS Random sampling

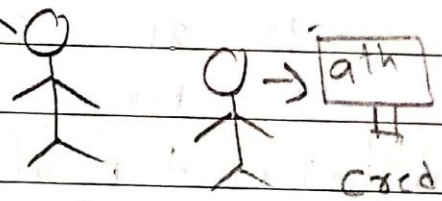
Date: / /

③ Systamatic Sampling:

select any n^{th} person out of population (N)

n^{th} person → {Air Port} → selling {credit card}

5th person



select every n^{th} individual out of population (N)

④ Convenience Sampling:

only those who are interested in the survey will only participate

{Data science survey → General AI survey}

① Survey Regarding New technology:

→ Convenience Sampling

② RBI survey ⇒ [women] → married

→ Statified + Random Sampling

③ credit card:

stratified + Random Sampling

MATRIKAS

Date / /

①Variabe: A variable is a property that can take any variables

Eg: age = 14 Variable
age = 25 Age = [24, 25, 26, 27, 28]
age = 100 ↓ collection

Two diffierent type of Variable:

① Quantitative Variable:

Measured numerically (mathematical operation)

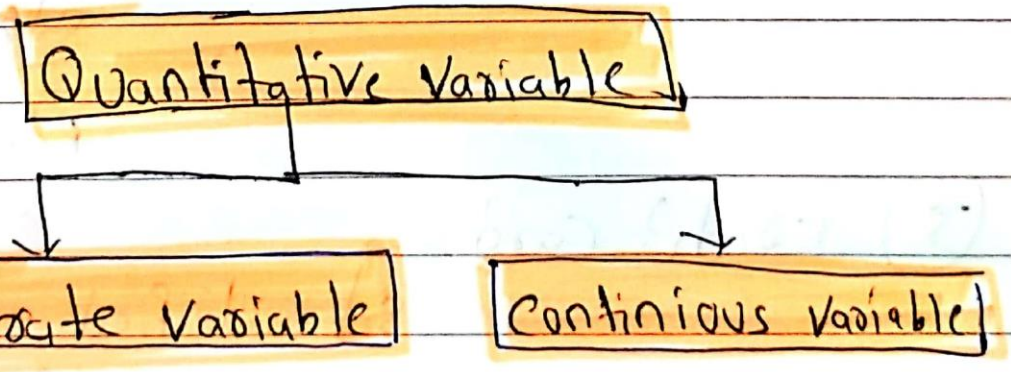
Ex: age, weight, height, dist, rainfall, temp,

② Qualitative Variable:

Categorical variables

(based on some characteristics they are grouped together

Eg: Gender, types of flowers, type of movies



MATRIKAS

Date: / /

Discrete Variable:

Eg:- Whole number → fixed

Eg:- N.o of Bank accounts

{1, 2, 3, 4} 2.5 ✗

Eg:- N.o of children :- whole n.o

Continious Variable:-

Eg:- continous → decimal values

Eg:- height, weight, ages, speed

marital → named
→ not named

→ Catagorical variable

Quantitative Variable

- Discrete Variable
- Continious Variable

Whole Numbers

Bank account = {1, 2, 1, 4}

Pin code = {.....}