Page No.
Date

* Small Assignment
* BIG Assignment

• Data Analyst Do ?

→ eg :- Amazon wants to decide when is the Next Big Billon Sale? month and with day ?

① Data Gathering

Tool → JIRA

↑ Business Analyst ↑ Product manager

↓

Domain knowledge

{SQL = Database}

② Data Analyst

→ ① what all data They require ??
→ ② Product manager (Domain Expert)
→ 3rd party, External Database.

Paid data.

↳ Statistical Analysis (Conclusion)

EDA [Exploratory Data Analysis] (understanding data)

Reporting [Power BI, Tableau]

↓ ↓

Stakeholders Key Performance Indicators [KPI-s]

↓

Decision making

Day 1

Page No.
Date

Statistics.

Def :-

Statistics is the science of Collecting, Organising and analysing data.

"Data is raw material".

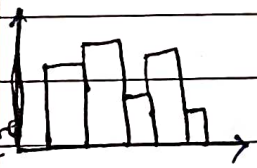
Eg :- Age of students in Classroom
{24, 30, 21, 34, 20} = Analysis.

Types of Statistics.

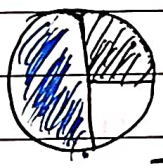
Descriptive Stats

* It is consists of Organising and summarising of Data.

eg :-



Histogram



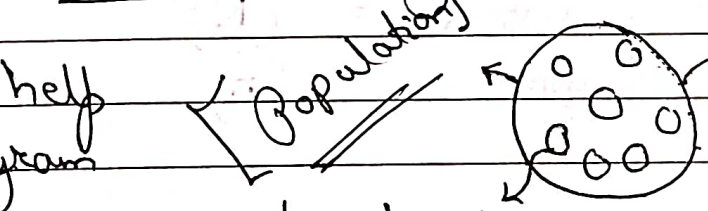
Pie Chart

Here with the help of Chart and diagram we are Organising the ~~data~~ data in Pie Or in Histogram.

Inferential Stats

* It consists of a Technique by Creating some measurment to form some Conclusion.

Eg :- ⇒ 18



As, here in west Bengal there is Election has been held over, ABP Tak make an Exit Poll To form a Conclusion, with the sample of Population of data Conclusion make.

Hypothesis Testing.
eg :- P value, Z test,

* Interview Question.

what is the average size of all sharks in the world ?

Central limit theorem

$n > 30 \approx$ Gaussian distribution

* Sampling Techniques.

(i) Simple Random Sampling.

Every member of population has an equal chance of being selected for your sample.

Population means entire data.
Population = (N)
Sample = (n)

(ii) Stratified Sampling

Stratified means Strata → Layers which are Non Overlapping

Page No.
Date

eg :- Gender → male, female
Education degree → BE, master, PHD, High school

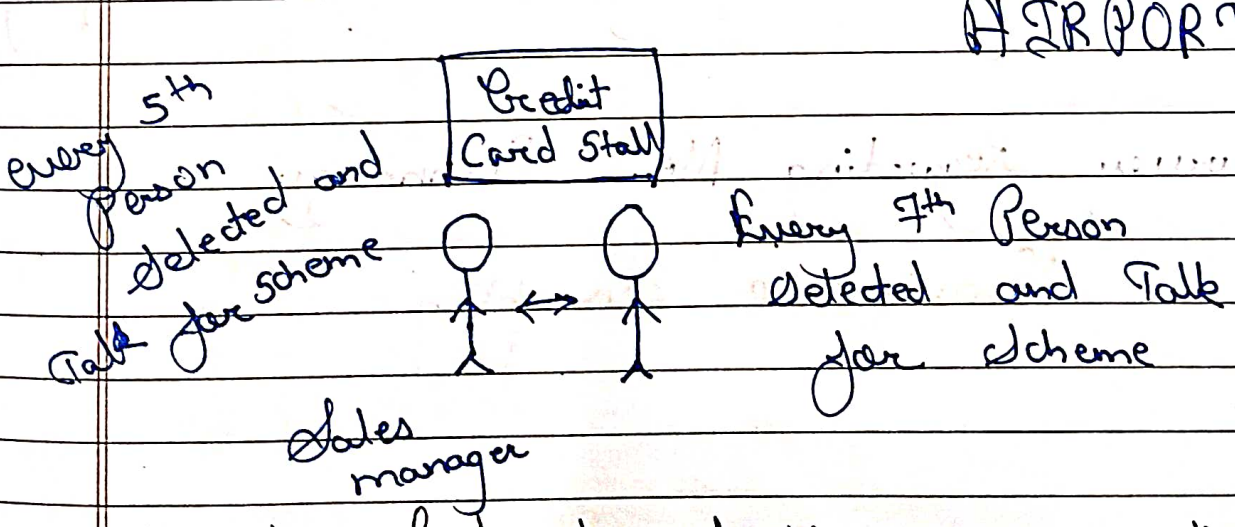
Here we can see it are completely seperate layer of group.

* As with education degree we sees an overlaps over layers.

Education → BE, masters, PHD, High school

Here after High school a student perseive BE after that masters, PHD so here is some overlaps.

(iii) Systematic Sampling



As in systematic sampling every n^{th} individual has been selected.

(iv) Convenience Sampling

- Only those people who are interested will only be participating

Data Science Survey → Arts
→ Data Science
→ medical.

Some Examples of Sampling

(1) Sales Credit Cards.

Systematic Sampling & Random Sampling

(2) Exit Polls (age > 18)

Stratified Sampling & Random Sampling

(3) Survey Regarding New Technology

Convenience Sampling

* Variables

A Variable is a property that can take on any value.

eg :- Age = 24
Age = 25
Age = 29

Two Kind of Variable.
① Quantitative Variable (measured Numerically)
② Qualitative Variable (Categorical Variables)
eg :- Gender { Male, Female } Based on Characteristic we can derive graph.

Quantitative Variable

Discrete Variable (whole number)

eg: No. of Children in family.
{0, 1, 2, 3, 4, 5}
2/5 Children
3/5 Children

Continuous Variables (Numerical Variable)
eg : Height = 172.5
106.62 cm

* Assessment

(1) what kind of Variable is marital status?

→ Categorical (married, widow, Unmarried)

(2) what kind of Variable is River Length?

→ Continuous Variable.

(3) what kind of Variable is movie duration?

→ Continuous Variable.

* Histograms {Construct a histogram}

Ages = {10, 12, 14, 18, 24, 26, 30, 35}
(10.5)
Continuous Values.

(1) Sort the Number {Asscending Order}
(2) Bins size {groups} or gaps.

Bin → no. of group.
Bin Size → Size of Bins

[0 - 100] Bin = 10

Bin Size = $\frac{100}{10}$ = 10

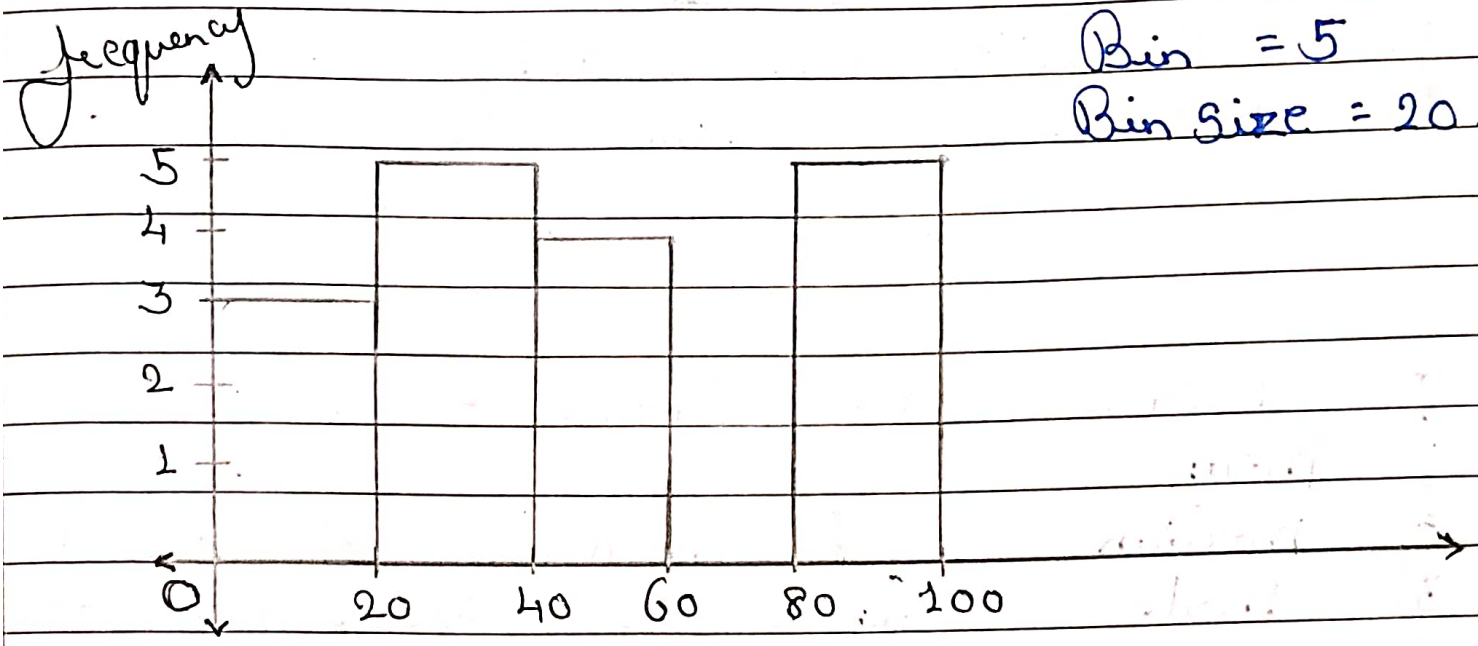
[0-100] Bin = 5

Bin Size = $\frac{100}{5}$ = 20

PdF (Probility density function) Smoothing histogram

Assessment

eg :- {10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99}



Bin = 5

Bin Size = 20.