

Statistic BI

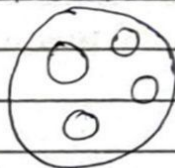
what is statistics and its types?

Statistics:- Concerns the collection, organization, analysis, interpretation and Presentation of Data

Descriptive Stats

- ① Analyse Data, Summarizing, organizing in form NUMBER or GRAPH
- ② Bar plot, Hist, Piechart, PDF, CDF
- ③ Measure of Central Tendency
Mean, Median, Mode
- ④ Measure of Variance
Std. dev, Varian, Range

Inferential Stats

- ① state of 7M
- 
- Exit Poll

Party 1 \rightarrow $x\%$

Party 2 \rightarrow $y\%$

Party 3 \rightarrow $z\%$

- ② — Sample \Rightarrow Inference + Conclu for this Population
 \Downarrow
for Population

③ Confidence intervals

E.g. Z-test, T-test
Chi-sq test

\leftarrow Hypo testin

Population vs Sample

Population:-

e.g. find Avg income of People living in Maharashtra.
 $1.2M = \text{Population}$

$$\text{Population Mean} = \frac{1}{1.2M} \sum_{i=1}^{1.2M} x_i$$

Here it is impossible to calculate for such huge Population.

Sample :- 10000 People from Maharashtra

$$\text{Sample mean} = \frac{1}{10000} \sum_{i=1}^{10000} x_i$$

Population Mean > Sample Mean.

Measure of Central Tendencies

① Mean = Average \Rightarrow Sum of values / No. of values
$$\text{Mean} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{--- for } n\text{-sampled}$$

② Median = Value that divides dataset in two halves
$$\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ term} \quad \text{--- } n = \text{odd}$$

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{\text{th}} + \left(\frac{n}{2} + 1\right)^{\text{th}}}{2} \text{ term} \quad \text{--- } n = \text{even}$$

③ Mode = Value with greatest frequency
Use in Categorical Variable.

• while treating Missing Values

	Variable Type	Outlier	Replace with
①	Numeric	Yes	Median
②	Numeric	No	Mean
③	char	Yes/No	mode

Measure of Variations

① Range: Difference betⁿ the Highest and Smallest element

② Variance:- Variance tells us how far data is spread

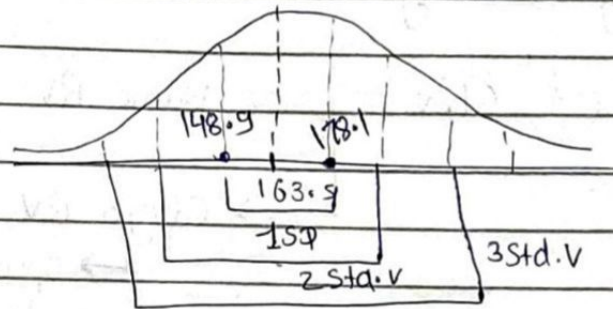
Sample of Height = { 168, 170, 150, 160, 182, 140, 175 }

$$V = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 = 213.95$$

→ mean = 163.57

③ Std. deviation:- Spread of Value within Distribution

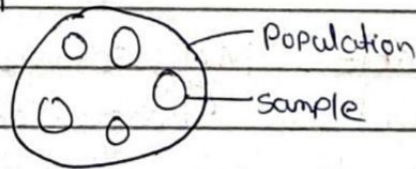
$$\sigma = \sqrt{V} = 163.5 \text{ (above ex)}$$



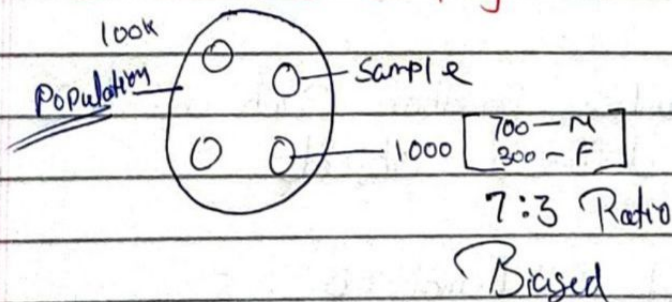
What are Diff. Sampling Technique

① Random Sampling Tech:-

Randomly select some Sample to infer some conclusion for Population



② Stratified Sampling Tech:-



In Stratified Sampling we will Pick Sample in 1:1 Ratio.

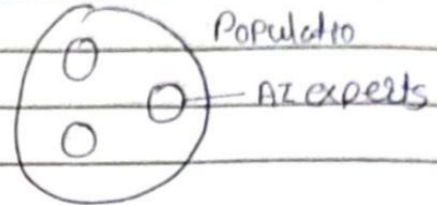
Due to this we may get improper inference

③ Systematic Sampling:-

Selecting every n^{th} element from sample Popul.

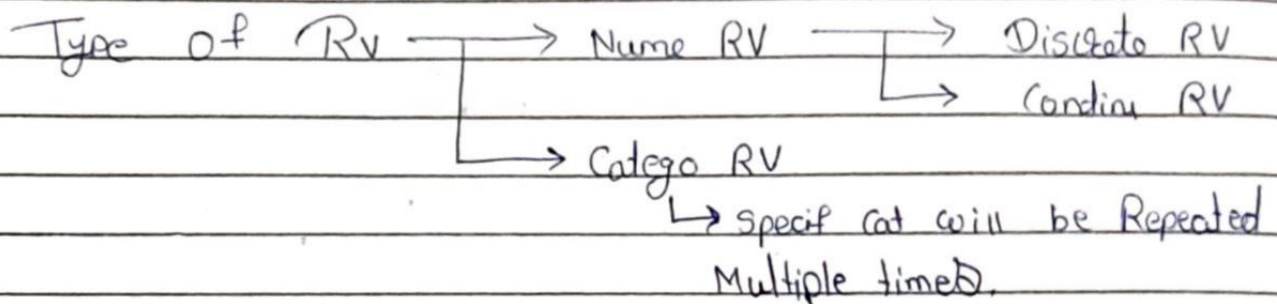
④ Cluster Sampling:- Domain Specific Selection

E.g. Survey on AI



What are Random Variables and its type

Random Variable is Place where we are storing something.



e.g. ① Discrete RV \Rightarrow No of People —

\hookrightarrow whole and Non-negative

② Continuous RV \Rightarrow Salary.

\hookrightarrow Any number

Variable Measurement Scales

Measurement is Process of applying number to object according to set of Rule.

* Scales of Measurements

① NOMINAL :-

Assigning Number so that each number

Representing different objects

e.g. Gender

Male = 1 , Female = 0

② ORDINAL:-

Assigning Number to object but here assigned number has its own meaning.

Ex.1. Ranks in class:- 1st, 2nd, 3rd ... etc.

③ Interval:-

Numbers have order but there are also equal intervals betⁿ adj category.

Ex.1 Percentage interval:- 5%-10%, 10%-15%

④ Ratio:- Ratio Scale is order scale in which the diff betⁿ the measurement is a meaningful quantity and measurement have a true zero Point.

Frequency Distribution

① Discrete Distribution

No. Students	Marks
4	20
2	40
5	30
6	62

2nd student = 40 marks

② Continuous Distribution

No. of student	marks
4	10-20
2	45-50
5	20-40
6	12-32

2nd student → 45-50 Class

Cumulative freq

Mark	No. of stu	C.F.	C
0-25	2	2	2
25-50	5	5+2	7
50-75	3	5+2+3	10
75-100	2	5+2+3+2	12

