19/01/25

# Stats

## Day-1

* Introduction to stats :-

1. Basic to Advance {Data Scientist, Data Analyst, Business Intellgent Tools}

→ First 2 days Basics

1. Descriptive Stats
2. Inferential Stats

1. Descriptive stats :-

{ i. Measure at Central Tendency }
{ ii. Measure of Dispersion }

Any thing relating to Summarizing data:
→ Histograms, pdf, cdf, probabilits, Permutation, mean, median, mode, varience, Standard deviation, ...

1) Gausian Distributin
2) LogNormal Distributin
3) Binomial Distributin
4) Bernoulis Distributin
5) parito Distributn {Power law Distr}
6) Standard Normal distributn
7) Transformation and Standardization
8) Q-Q plot

2. Inferential stats :-

Z-test
T-test
Anova
Chisquare
Hypothesis -testing

Z, table , T table .

Now,

Q. What is Statistics ?

→ Stastistics is the Science of Collecting organizing and analyzing data.
{Better Decision making

Definition of data ?

→ Facts or pieces of informatin that Can be Measures.

Eg:- i) The IQ of a class a student
{98, 97, 60, 55, 75, 65}

ii) Ages of Students of a Class
{30, 25, 24, 23, 27, 28} → Data

* Types of Statistics :-

i) Descriptive Stats.
→ It Consist of Organizing and Summarizing data.

ii) Infrential Stats.
→ It is a Technique where we use the data that we measured to form Conclusion.

Eg:- Classroom of maths Student, (20 Student)
Marks of the 1st Sem

84, 86, 78, 72, 75, 65, 80, 81, 92, 95, 96, 97. - - - - - .

Eg:- Descriptive Stait:-

What is the average marks of the Studens in the class.

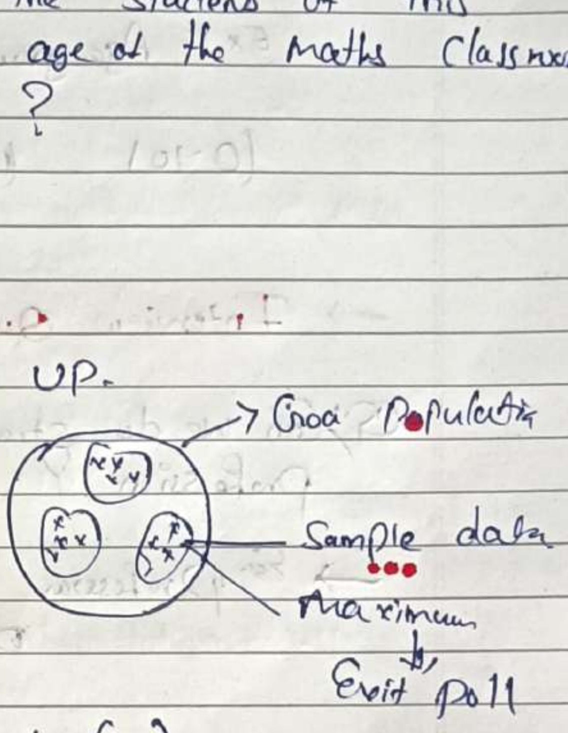
Eg:- Infrential Stats:-

→ Are the marks of the Studens of this Classroom similar to the age of the maths Classroom in the college. ?

Population and Sample:-

Elections → Goa, UP.

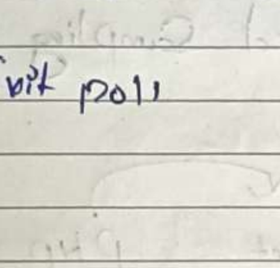
Exit poll



Populatica (N)  Sample (n)

* Sampling Techniques :-

i) Simple Random Sampling :-



When Performing simple random Sampling Every nb. of populatin (N) has an Equal Chance of being Selected for your sample N.

## Strat :- layering

ii) Stratified Sampling :- is a technique where the population is split into Non-overlapping groups (Strata)

Eg :- Gender → Male
→ Female
Survey.

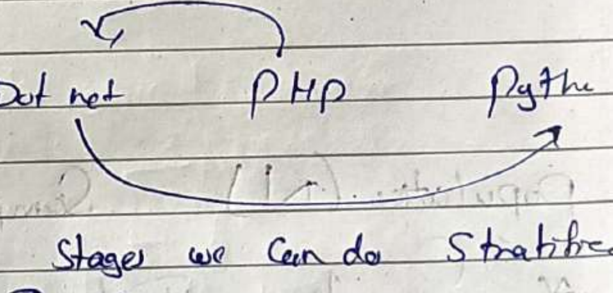
Ex :- Age-group

[0-10) [10-20) [20-40) [40-100)

→ Interview Question :-

• Q) Can we do stratified Sampling based on profession ?

→ Eg. Profession



⇒ Some of the stages we can do Stratified Sampling

→ But, applying stratified sum other Conditions we can make sure that the Sampling satisfies.

iii) Systematic Sampling :-

(N) → $n^{th}$ individual

Eg:.
mall → Survey (covid)

↳ 8th person → Survey
↳ 10th person → Survey

(Indipendantly Survey)

iv) Conveince Sampling :-

Eg:- ↳ Survey
↳ Data Science ← Only these people

{Basically intrest / Expert in DS}

→ Survey related to Specific topic.
→ In this case Data Science)

Eg:- Exit poll
{Random Sampling using}

Eg:- RBI → House hold Survey
⇓
Survey → women
{Stratified Sampling / Conveince Sampling}

Eg:. Drugs → Tested ⇒ . . .

*** Variables** :- A variable is a property that can take on any value

Eg :-

Height {182, 178, 168, 180, 172}
weight {78, 99, 100, 60, 72}

Two kinds of variable :-

1) Quantitative variable
11) Qualitative variable / Categorical variable

⇒ 1) Quantitative Variable :-

Measured Numerically {Add, Substract, mul, div}

Eg :- Age
weight
Height

11) Qualitative / Categorical variable :

Eg :- Gender — M
F

→ Based on some characteristic we can divide categorical variable.

Eg :- IQ

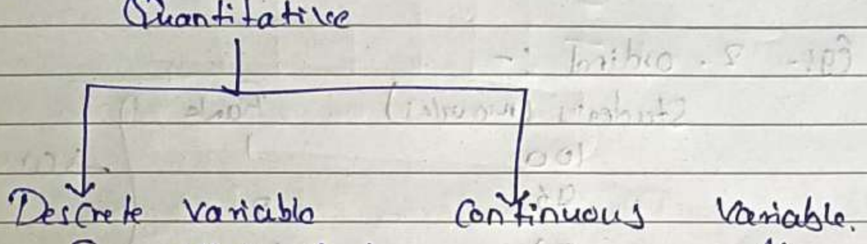
| | | |
|---|---|---|
| 0-10 | 10-50 | 50-100 |
| ⇓ | ⇓ | ⇓ |
| Less IQ | Medium IQ | high IQ<br>Good IQ |

Page No.
Date

Eg:

| Blood group | Tshirt |
|---|---|
| A + | xs |
| A - | 5 |
| O+ | m |
| O+ | L |
| Etc | xL |
| | xxL, etc |

⇒



Quantitative

Descrete variable

Continuous Variable.

Eg - Whole numbers

Eg:- No of Bank Acc
→ 2, 3, 4, 5, 6, 7

Eg. Height
172.5, 162.2, 172.3, Etc

Eg:
→ No. of childrens in a family
Eg: 3, 4, 5, 2, 3, 4.

weight - 100, 99.5, 77.5, Etc.

Amount of Rainfall:
1.1, 2.3, 1.35, Etc.

Sample Questions:-

1. What kind of Variable Gend is? ⇒ Categorical.
2. " " " " marital Status? ⇒ "
3. " " " " Population of State is? ⇒ Discre
4. " " " " River length? ⇒ Continuous
5. " " " " Song length? ⇒ Continuous
6. " " " " Blood pressure? ⇒ ~~Discrete~~ Continue
7. " " " PIN code? Discrete

* Variable Measurement Scales :-

4. types of measure of variable

1. Nominal → {Categorical data} → Classes Eg:- Color, Gender
2. Ordinal → Order of the data matters, value does not
3. Interval → order matters, value matters, natural zero not Present
4. Ratio.

Eg:- 2. ordinal :-

| Students (marks) | Rank |
|---|---|
| 100 | 1 |
| 96 | 2 |
| 57 | 4 |
| 85 | 3 |
| 44 | 5 |

→ ordinal data

3. Interval :- Order matters, value also matters, natural zero not Present

Eg:- Temparature
→ Fahrenheit
70-80 80-90 90-100 100-110

4. Ratio :-
Absolute zero points, Equal intervals, Ratio are meaningful, Quantitative

eg:- length (measure in meters, cm, etc)
Weight (" in kg, grams etc)
Time (" " Second, min, hrs, etc)
Income . (" " Currency, dollars, etc)
Age. (" " years, months, days, dy

* Frequency Distribution :-

Sample datset :- Rose, lilly, Sunflower, Rose, lilly, Sunflowers, Rose, lilly, lilly

| Flower | Frequeny |
|---|---|
| Rose | 3 |
| lilly | 4 |
| Sun flower | 2 |

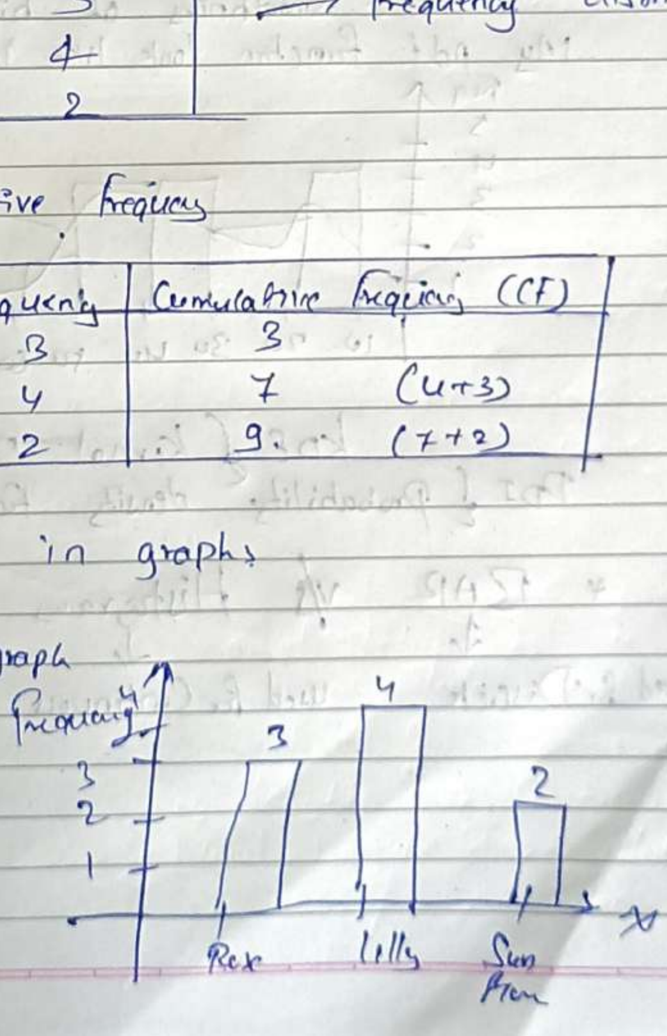
→ Frequency distribution table

-> Cumulative frequecy

| Flower | Frequency | Cumulative frequency (CF) |
|---|---|---|
| Rose | 3 | 3 |
| lilly | 4 | 7 (4+3) |
| Sunflower | 2 | 9. (7+2) |

Eg;- in graphs

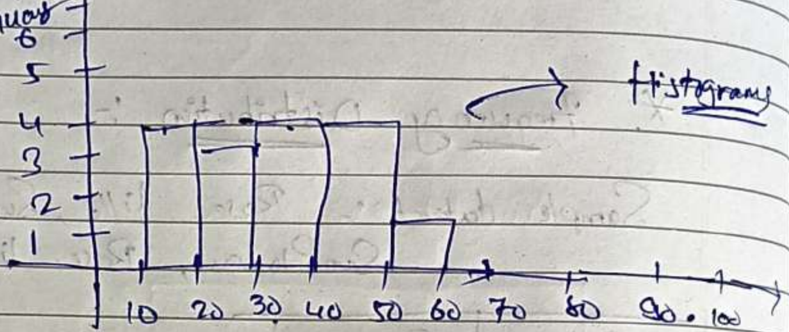
i) Bar graph



Eg:- 11) Histogram :- Continuous :-

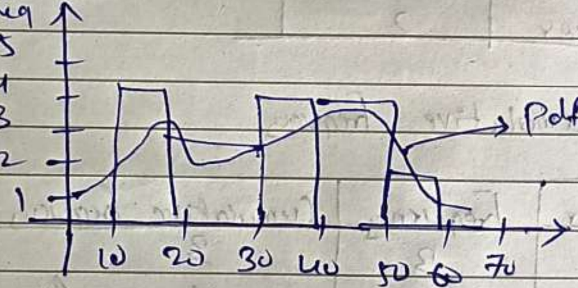
Age = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Bin = 10 ⟹
↓
(Grouping)



Pdf :- Smoothning of histogram

My pdf function look like ↓



KDE {Kernal Density Estimator}

PDF {Probability density function}

* BAR V/S Histograms

↓ ↓

Used for Discrete    Used for Continuous