Lecture-01

STATISTICS

Introduction to Stats:

Statistics is the branch of applied mathematics that involves the collection, description, analysis and inference of Conclusions from quantiative data.

Statistics is the Science of the data to Collect, Organize and analysis to made Conclusions or decision Making.

Descriptive Stats	Inferential Stats
It focus on Collecting, Organizing and Summarizing the data.	It focus on to made Conclusion. It is the technique where in we Used the data that we have Measured to form Conclusion.

Descriptive Stats:

↓

1. Measure of Central tendency.

2. Measure of Dispersion .

Summarizing the data,

Histograms , pdf, cdf, Probability, Permutation , Mean, median, mode , Varience , standard deviation.

Distributions,

1. Gaussian (or) Normal distribution .

2. Log Normal Distribution .

3. Binomial Distribution .

4. Bernauli's distribution .

5. Poisson distribution .

6. Pareto distribution .

7. Transformation and Standardization .

8. Q -Q Plot .

Inferential Stats:

↓

(F-Test)

Z test, T test , ANOVA Test , CHISQUARE,

Hypothesis test (P values) , Confidence Interval,

Z-table , t-table .

What is Data ?

Facts or pieces of information that can be measured.

Data

Eg: IQ of students = {90, 80, 50, 100 , 120}

Student's marks = [90, 80, 85, 70, 60, 30, 55, 85, 93]

↳ one particular math classroom marks

The Common question from above Marks in the Point of descriptive stats are,

1. What is the average students marks ?

2. How many students are get marks above 50 ?

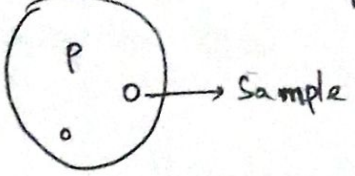
etc.

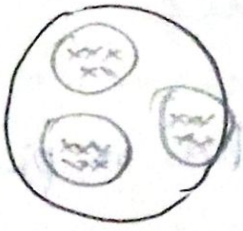
The Common question from above marks in the Point of inferential stats,

1. Are the marks of the students of this Classroom Similar to the number of Maths Classroom in the College?

Whole math Classroom

Particular one math Classroom





Population (N) - Entire Population among the data.

Sample (n) - Some Sample from entire population of the data.

<u>Same Sampling Technique:</u>

<u>Simple Random Sampling:</u>

Every number of the population (N) has an equal Chance of being Selected for your Sample (n).

It focus on random Sampling.

Stratified Sampling:

Where the population (N) is split into non-Overlapping groups. (strata)

Ex: we want to collect and sample about Job professions. we need to sample get from only Job profession with respective field doctor, engineer, IAS etc. [Not applicable for collect sample from whole population]

Ex 2: Gender → male
→ female

Non-overlapping groups - One group independent from another group.

Systematic Sampling:

(N) → nth individual

Eg: Mall - Survey about Covid for every 10 th person out from the mall.

Systematic Sampling depends on the nature or Scenario.

Convenience Sampling:

Samples are collected from population if only for Convenience.

Example: DATA SCIENCE
↑
only get Samples from DATA SCIENCE in Convenient way.

Examples for Sampling:

VOTE POLL SAMPLE	- Random Sampling used.
RBI Women Survey	- Convenience Sampling.
DRUG TESTED	- Based on Condition and nature.

Variables:

A Variable is a property that can take on any value. [STATS DEFINITION]

A Variable is the Container to store the Value. [PROGRAMMING DEFINITION] Both are Same

Eg:

Height = {78, 65, 60, 40, 50}

weight = {65, 80, 100, 120, 75}

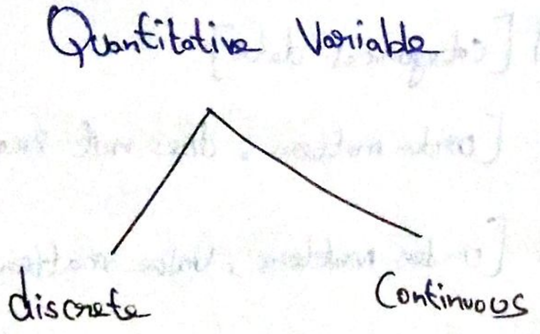
Two kinds of Variables:

→ Quantitative Variable.

→ Qualitative Variable.

Quantitative Variable:

It can be measured and numerically type and supports Operation for Add, mul etc..



discrete	Continuous
Eg: Whole number [0,9,10]	Eg: height: [172.5, 162.5 ..
No of Bank accant -5	weight: [100kg, 99.5]
[Cannot be split]	[Can split]

Qualitative Variable / Categorial Variable:

It is based on the data other from Quantitative like strings.

Gender → f
Gender → m

{Based on Some Characteristics we can derive Categorical Variable}

Example:

Gender, email etc...

Variable Measurement Scales:

4 types of measured Variable.

→ Nominal [categorical data]

→ Ordinal [order matters, does not require Values]

→ Interval [order matters, value matters, natural 0 are not Present]

→ Ratio [order matters, Value Matters, Natural 0 present]

Nominal - Male.

Ordinal -

RANK	marks
1	95
2	80
3	70

→ This ordered data is known as Ordinal.

Interval - $\boxed{70-80}$ $\boxed{80-100}$ but 0 not present

50° 60°

0 → does not respect .

Ratio - It possesses all the properties of interval data and allows for meaningful ratios between Values.

height : [180 , 180, 170]

$$\begin{array}{ccc} \downarrow & \downarrow & \downarrow \\ A & B & C \end{array}$$

The ratio of person B height to person C height

$$= \frac{180}{170} = 18:17 \text{ (or) } \simeq 1.05883$$

FREQUENCY DISTRIBUTION :

→ It is the distribution to return and Visualize the Count through Various graphs like bar graph, Histogram .

→ Let Consider the Simple example,

Number of population in Countries like .

Sample dataset:

Country = [India, nepal, nepal, india, india, india,
Yeman, Oman, Oman].

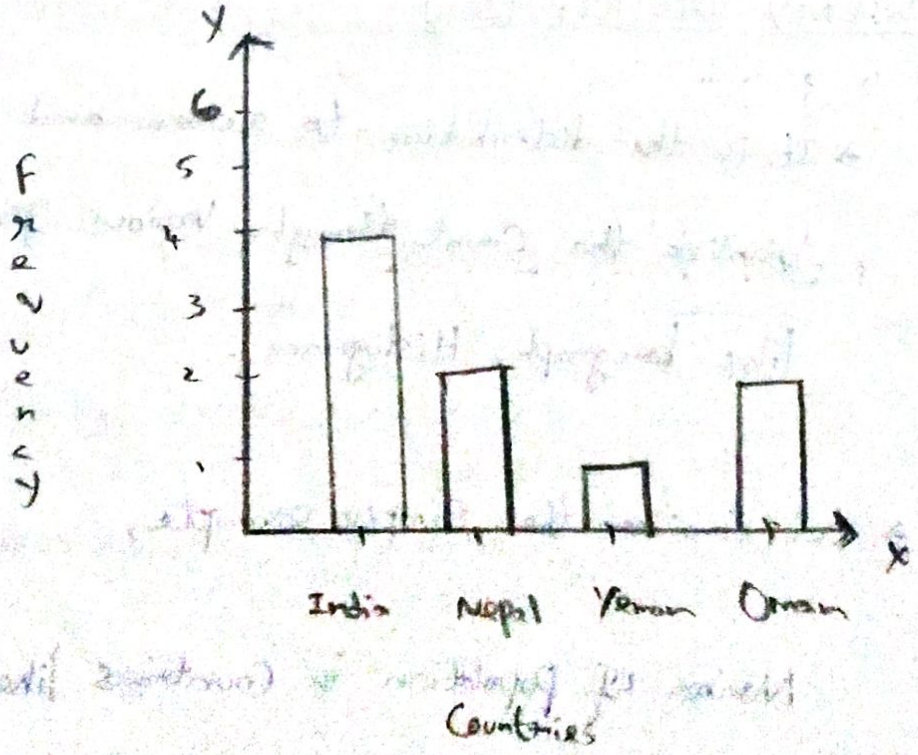
Countries	frequency	Cumulative frequency
India	4	4
Nepal	2	6
Yeman	1	7
Oman	2	9

↑ Total Count

Visualize through,

(Summarizing the data)

BAR GRAPH:



Bar graph Only represents the discrete data (or)

Categorical data.

Suppose Use Continuous to use histogram,

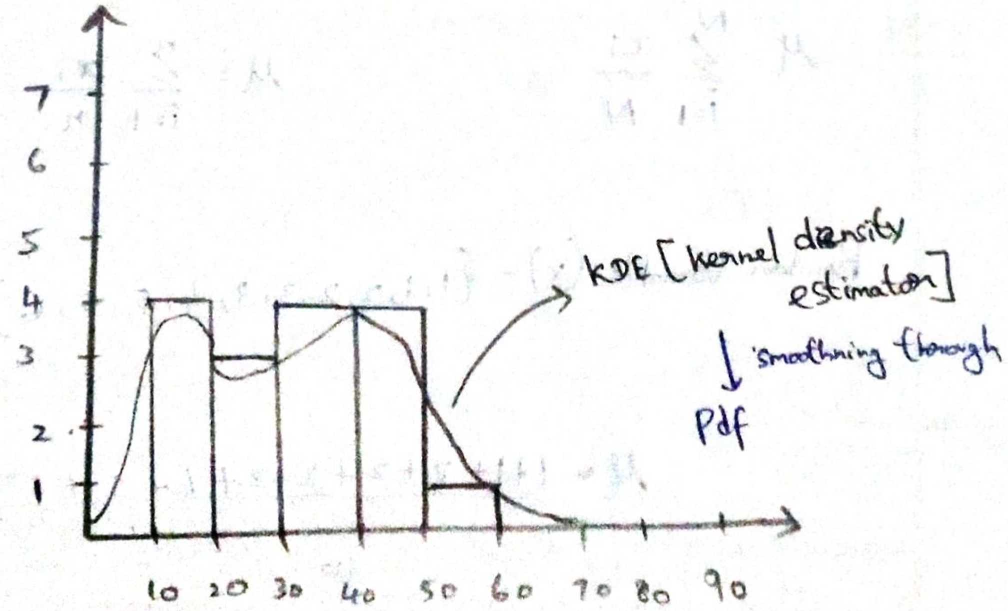
HISTOGRAM:

Sample data,

age = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41,

42, 43, 50, 51}

Default histogram bin Size = 10 [able to change]



Pdf is nothing but Smoothening of histogram.

Pdf [probability density function]