Machine Learning

Unit - I

Statistics :- Statistics is the branch of applied Mathematics, which specialises in data.

Statistics is the methodology which scientists and Mathematicians have developed for interpreting and drawing conclusions from collected data.

Statistics Method is also called scientific Method.

or

Statistics consists of a body of Methods for collecting and analysing data.

Population :→ Population is the collection of all individuals or items under consideration in a statistical study.

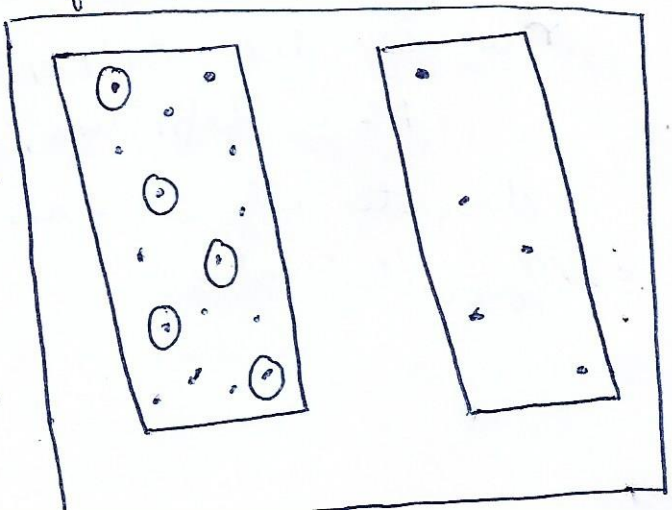
or

it is the set of Measurements (or Record of some Qualitative Traits) Corresponding to the entire collection of units for which Inferences are to be made

Sample :- Sample is that Part of the Population from which Information is collected.

or

A Sample from statistical Population is the set of Measurements That are actually collected in the course of an investigation.



Variables and organizations of the data :→

A characteristics that varies from one person or things to another person or things is called a variable i.e. a variable is any characteristi that varies from one individual member of the population to another.

Example of variables for humans are height, weight, no. of siblings, sex, marital status, and eye color.

The first three of these variables are the examples of quantitative (or numerical) variables, and the last three are example of qualitative (or categorical) variables

Quantitative variables can be classified as either discrete or continuous variables.

Discrete variables:- Some variables, such as the no. of children in family, the no. of car accident on the certain road on different days.

typically, a discrete variable is a variable whose possible value are some or all of the ordinary counting number like 0, 1, 2, 3...

As a definition, we can say that a variable is discrete if it has only a countable no. of distinct possible value.

Continuous variable:- quantities such as lengths, weight or temperature can ~~in principle~~ be continuous variable. A continuous variable is a variable whose value is obtained by measuring, ie one which can take on an uncountable set of values. For Example, a variable over a non empty range of real numbers is continuous. If it can take on any value in that range.

Frequency:- Relative frequency of the class

$$= \frac{\text{frequency in the class}}{\text{total no. of observation}}$$

The Aim is to focus on certain features of the data, which will describe their nature in a general way. The most important features are

(i) Central tendency (ii) Dispersion

(1) Measures of central tendency :-

There are some central value around, which other points are actually clusters. This will be represented from table, where the data seems to cluster around 1300 gm

Item No	Yield (gm)
1	1216
2	1374
3	1167
4	1232
5	1407
6	1453

M/s of central of tendency also known as M/s of location.

Definition:- There is a tendency of data to cluster around some central value and the M/s of this central value is called M/s of central tendency.

Method of calculation central tendency

1. Arithmetic Mean (AM)
2. Median
3. Mode
4. Geometric Mean (GM)
5. Harmonic Mean

} These all are Averages depending upon Particular condition

Arithmetic Mean :- AM of a set of observation is their sum divided by no. of observation e.g. of AM in Discrete case :

60	70	80	40	50
x_1	x_2	x_3	x_4	x_5

Different realisation of variable x

Let $x_1, x_2, x_3 \dots x_n$ are the different value of variable x.

The AM of x is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where x_i is summation of all observation and n is no of observation.

in ~~continuous~~ Discrete case :- when Frequency are given

eg.

F	5	2	7	6	1
X	32	25	40	50	58

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i$$

where $N = \sum_{i=1}^{n} f_i$ sum of frequency

$$\bar{x} = \frac{160 + 50 + 280 + 300 + 58}{21}$$

$$= 40.38$$

Ex:-

x :	1	2	3	4	5	6	7
f :	5	9	12	17	14	10	6
fx :	5	18	36	68	70	60	42

$N = \Sigma f = 73$

$\Sigma f_i x_i = 299$

$\bar{x} = \frac{1}{N} \Sigma f_i x_i = \frac{1}{73} \times 299$

$= 4.09$ Ans

Combined Mean:- Let there be Two sets of value

Let x_1 : The value in 1st set

x_2 : The value in 2nd set

$\bar{x}_1$ = AM of 1st set

$\bar{x}_2$ = AM of 2nd set

The Combined Mean of both the sets

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Ex:- Let The No. of male emp be 80 with Avg salary Rs 5200 and The No. of female employee be 20 with Avg salary Rs 4200. find The combined Mean.

$n_1 = 80, n_2 = 20$

$\bar{x}_1 = 5200, \bar{x}_2 = 4200$

$$\bar{x} = \frac{5200 \times 80 + 4200 \times 20}{80 + 20}$$

$\boxed{\bar{x} = 5000}$ Ans

Median :- Median is the middle most value of the observation, when data is arranged either in Ascending or in descending order i.e. Median divides the frequency distribution into 2 equal halves. Such that half of the values are below the Median and half of the values are above Median

(i) Discrete case :-

(1) when frequency are not giving

(a) when n is odd, Median from the same data set

$$\text{Median} = \left(\frac{n+1}{2}\right)^{Th} \text{ of the value}$$

eg 69, 59, 70, 82, 50

let say ascending order

50, 59, 69, 70, 82

$n = 5$

$$\text{Median} = \frac{5+1}{2} = \frac{6}{2} = 3$$

3rd ordered value of arrange data = 69

(b) when n is even

$$\text{Median} = \text{Mean of } \left(\frac{n}{2}\right)^{Th} \text{ and } \left(\frac{n}{2}+1\right)^{Th} \text{ value}$$

69, 59, 70, 82, 50, 90

in Ascending order

50, 59, 69, 70, 82, 90

$\frac{n^{th}}{2}$ value $= \frac{6}{2} = 3^{rd}$ value $= 69$

$\left(\frac{n}{2}+1\right)^{th} = \left(\frac{6}{2}+1\right) = 4^{th}$ value $= 70$

Median $= \frac{1}{2}(69+70) = 69.5$

* Linear combination of two Median will be Median

(ii) <u>when frequency are given</u> :- In case of discrete ~~in case of~~ frequency distribution, median is defined/ obtained by considering the commutative frequency.

The steps are

(i) find $\frac{N}{2}$ where $N = \sum_{i=1}^{n} f_i$

(ii) see the commutative frequency just greater than $N/2$

(iii) The corresponding value of x is Median

<u>Ex:-</u> let

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6
less than CF	8	18	29	45	65	90	105	114	120

(i) $\frac{N}{2} = 60$ ← more than

(ii) 65, Median is the value of x corresponding to 65 is 5

* If 60 is present in CF then we have to choose that, instead of 65.

Drawback of AM :—

*

10, 10, 15, 15, 150

$$\bar{x} = \frac{100}{5} = 20$$

Here, four out of five observation are less than AM, this is because of extream value 50, AM is affected by the such value (50) which is away from the bunch of data. Therefore AM will be affected in case of extream value.

but Median does not affected.

Mode :— The mode of a distribution is that value of variable, which occur most frequently in the set of observation or the value of variable with highest frequency; Discrete case :— when frequency is not given

eg 20 22 22 28 80

mode : 22

* least favorable method than Mean and Median, mode is not unique it may be more than once.

Ex :— 20 22 22 28 30 30 28

* unimodal distribution
* bimodal distribution
* multimodal distribution

10, 15, 20, 25 ; each no. working as mode

Discrete case:- when frequency is ~~not~~ given

Ex:-

x:	1	2	3	4	5	6	7
f:	10	15	25	20	18	7	8

The value of x corresponding to the max freq 25 is 3. So here mode is 3.

* Mode = 3 Median − 2 Mean

Ex:- For this distribution Mean is found to be $\bar{x} = 164.734$ cm and the Median is found to be $M_i = 164.758$ cm.

Therefore $M_o = 3 \times 164.758 - 2 \times 164.734$

$\Rightarrow 164.806$ cm (approx).

Properties of Mode:-

* Mode is not at all affected by extream values.

* It is not based upon all the observations