INTRODUCTION TO STATISTICS & DATA COLLECTION

Statistics:

Statistics is the " *science which deals with the collection, analysis and interpretation of numerical data*".

Main Functions of Statistics:

- Collection of Data
- Presentation of Data
- Analysis of Data
- Interpretation of results

Types of Statistical Data:

- **Primary Data**: Primary data are those which are collected from the units or individuals directly and these data have never been used for any purpose earlier.

- **Secondary Data:** The data, which had been collected by some individual or agency and statistically treated to draw certain conclusions and now the same data are used and analyzed to extract some other information.

- Population
- Sample
- Parameter
- Statistic
- Sampling
- Random Sampling
- Non-Random Sampling

- Variable
- Frequency
- Discrete frequency distribution
- Continuous frequency distribution

Marks	20	30	40	50	60	70
No. of Students	8	12	20	10	6	4

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	12	18	27	20	17	6

Marks	0-9	10-19	20-29	30-39	40-49	50-59
No. of Students	12	18	27	20	17	6

Formation of Frequency Distribution

Classification according to class-intervals:

(i) Class Limits

(ii) Class-interval

(iii) Class-frequency

(iv) Class Mid-point

(v) Exclusive method

(vi) Inclusive method

Data Analysis:

- Measures of Central tendency or average
- Measures of Variation or dispersion
- Measures of Skewness
- Measures of Kurtosis

Measures of Central tendency or average

- Arithmetic Mean
- Median
- Mode
- Geometric Mean
- Harmonic Mean

Arithmetic Mean (A.M.)

A.M. : Arithmetic mean of set of observations is their sum divided by the number of observations.

Simple A.M. : $\bar{X} = \frac{x_1 + x_2 + \ldots\ldots\ldots + x_n}{N} = \frac{\sum_{i=1}^{n} x_i}{N}$

N - Number of observations.

Ex: A Monthly income of 10 families in a city is given by:

1600, 1560, 1440, 1530, 1670, 1860, 1750, 1910, 1490, 1800.

$$\bar{X} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{16610}{10} = \mathbf{1661}$$

A.M. For Discrete Series:

Direct Method: $\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$

f - Frequency of the given set of observations

$N = \sum_{i=1}^{n} f_i$ = Total number of observations

Ex: The following data represents the marks obtained by 60 students of a class. Obtain the average marks.

Marks	20	30	40	50	60	70
No. of Students	8	12	20	10	6	4

Marks	No. of Students (f)	fx
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
	$N = 60$	$\sum fx = 2460$

$$\bar{X} = \frac{\sum f_i x_i}{N} = \frac{2460}{60} = \mathbf{41}$$

A.M. For Continuous Series:

Direct Method: $\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$

f - Frequency of the given set of observations

x - mid-point of each class

$N = \sum_{i=1}^{n} f_i$ = Total number of observations

Ex: Obtain the Arithmetic mean for the following data.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	12	18	27	20	17	6

Marks	No. of Students (f)	Mid-point (x)	$f\,x$
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330

$$\bar{X} = \frac{\sum f_i x_i}{n} = \frac{2800}{100} = 28$$

Using the Deviation:

If the values of ***x*** or ***f*** are large, the calculation of A.M. by above formula is quite time-consuming and tedious.

The A.M. is reduced to a great extent by taking the deviations

of the given values from any arbitrary point 'A' :

Let $d_i = x_i - A \Rightarrow f_i d_i = f_i (x_i - A)$

$$\bar{X} = A + \frac{1}{N} \sum_{i=1}^{N} f_i d_i \qquad or \qquad \bar{X} = A + \frac{h}{N} \sum_{i=1}^{N} f_i d_i$$

h or i – common magnitude of class

Ex:

C.I.	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	7	16	24	15	7

C.I.	**Mid-Value**	**Frequency (*f*)**	$\boldsymbol{d = \frac{x - A}{h}}$	***fd***
0-8	4	8	-3	-24
8-16	12	7	-2	-14
16-24	20	16	-1	-16
24-32	28	24	0	0
32-40	36	15	1	15
40-48	44	7	2	14
Total		77		-25

$$\bar{X} = A + \frac{h}{N} \sum_{i=1}^{n} f_i d_i = 28 + \frac{8 \text{ x } (-25)}{77} = 25.404$$

Median:

Median of a distribution is the value of the variable which divides it into two equal parts.

It is the value which exceeds and is exceeded by the same number of observations. Thus the median is called as a *"**positional average**"*.

Evaluation of Median: For ungrouped data,

(i) Odd number of observations.

(ii) Even number of observations.

For Grouped data:

(i) Discrete Frequency distribution:

a) Find $\frac{N+1}{2}$, where N – Total Frequency = $\sum_{i=1}^{N} f_i$.

b) See the (less than) cumulative frequency (c.f.) just greater than $\frac{N+1}{2}$.

c) The corresponding value of x is median.

Ex:

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6

Calculate the Median of the distribution.

x	f	$c.f.$
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	10
8	9	114
9	6	120
	N = 120	

$$\frac{N+1}{2} = 60.5$$

The cumulative frequency (c.f.) just greater than $\frac{N+1}{2}$ is 65 and value corresponding to 65 is 5.

$\therefore$ $\boldsymbol{Median\ is\ 5}$.

(ii) Continuous Frequency Distribution:

In case of continuous frequency distribution, the class corresponding to the *c.f.* just greater than $\frac{N}{2}$ is called the *median class* and the value of median is obtained by the following formula:

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

where l - is the lower limit of the median class,

f - is the frequency of the median class,

h - is the magnitude of the median class,

c - is the *c.f.* of the class preceding the median class

Ex: Find the median of the following data:

Wages (in Rs.)	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
No. of workers	3	5	20	10	5

Solution:

Wages (in Rs.)	***No. of Employees***	***c.f.***
2000-3000	3	3
3000-4000	5	8
4000-5000	20	28
5000-6000	10	38
6000-7000	5	43
	N = 43	

$$\frac{N}{2} = \frac{43}{2} = 21.5$$

Cumulative frequency just greater than 21.5 is 28 and the corresponding class is 4000-5000.

Thus the median class is 4000-5000.

$l = 4000$; $h = 1000$; $f = 20$; $c = 8$

$$\text{Median} = 4000 + \frac{1000}{20}(21.5 - 8)$$

$\therefore$ ***Median =*** **4675**.

Quartiles

First Quartile (Q_1) = Size of $\frac{N+1}{4}\ th$ item (Discrete series)

Q_1 = Size of $\frac{N}{4}\ th$ item (Continuous series)

$$Q_1 = l + \frac{\frac{N}{4} - c.f.}{f} \times i$$

Third Quartile (Q_3) = Size of $3\left(\frac{N+1}{4}\right)$ th item (Discrete series)

Q_3 = Size of $\frac{3N}{4}$ th item. (Continuous series)

$$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

Example 3: Calculate Q_1 and Q_3 for the following data.

Roll No.	1	2	3	4	5	6	7
Marks	20	28	40	12	30	15	50

Solution: Marks in ascending order 12 15 20 28 30 40 50

Q_1 = Size of $\frac{N+1}{4}\ th$ item = Size of $\frac{7+1}{4}$ = 2nd item.

Size of 2nd item is 15. Hence $\boldsymbol{Q_1 = 15}$

Q_3 = Size of $3\left(\frac{N+1}{4}\right)\ th$ item = Size of $3\left(\frac{7+1}{4}\right)$= 6th item.

Size of 6th item is 40. Hence $\boldsymbol{Q_3 = 40.}$

.

Example 4: Compute the value of Q_1 and Q_3 for following data:

C.I.	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	12	19	5	10	9	6	6

Solution:

Marks	*Frequency*	*Cumulative Frequency*
10-20	12	12
20-30	19	31
30-40	5	36
40-50	10	46
50-60	9	55
60-70	6	61
70-80	6	67
	$\boldsymbol{N} = 67$	

Q_1 = Size of $\frac{N}{4}\ th$ item = Size of $\frac{67}{4}$ = 16.75^{th} item.

Q_1 - lies in the interval **20-30**

$$Q_1 = l + \frac{\frac{N}{4} - c.f.}{f} \times i$$

l = 20, N/4 = 16.75, c.f. = 12

f = 19, i = 10

$$Q_1 = 20 + \frac{\frac{67}{4} - 12}{19} \times 10 = 20 + 2.5 = 22.5$$

Hence **Q_1 = 22.5**

Q_3 = Size of $\frac{3N}{4}\ th$ item = Size of $\frac{3\times67}{4}$ = 50.25th item.

Q_3 - lies in the class **50-60**.

$$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$l = 50, 3N/4 = 50.25, c.f. = 46$

$f = 9,\ i = 10$

$$Q_3 = 50 + \frac{50.25-46}{9} \times 10 = 50+4.72 = 54.72$$

Hence **Q_3 = 54.72**

Deciles: $\boldsymbol{D_4} =$ Size of $4\left(\frac{N+1}{10}\right)$ item in individual and discrete series.

$\boldsymbol{D_4}$ = Size of $\frac{4N}{10}$ th item in continuous series.

Percentiles: $\boldsymbol{P_{60}} =$ Size of $60\left(\frac{N+1}{100}\right)$ th item in discrete series.

$\boldsymbol{P_{60}}$= Size of $\frac{60N}{100}$ th item in continuous series.

$\boldsymbol{D_4}$ = Size of $\frac{4N}{10}$ th item = $\frac{4\times67}{10}$ = 26.8 th item.

$\boldsymbol{D_4}$ lies in the interval of 20-30.

$$\boldsymbol{D_4} = l + \frac{\frac{4N}{10} - c.f.}{f} \times i$$

$l = 20$, $4N/10 = 26.8$, $c.f. = 12$

$f = 19$, $i = 10$

$\boldsymbol{D_4}$ = 27.79

$\boldsymbol{P_{60}}$= Size of $\frac{60N}{100}$ th item = $\frac{60\times67}{100}$ = 40.2 th item

$\boldsymbol{P_{60}}$ lies in the interval of 40-50.

$$\boldsymbol{P_{60}} = l + \frac{\frac{60N}{100} - c.f.}{f} \times i$$

$l = 40$, $4N/10 = 40.2$, $c.f. = 36$

$f = 10$, $i = 10$

$\boldsymbol{P_{60}}$ = 44.2

MODE

Definition: Mode is the value of the variable which is predominant in the series.

1. For a discrete frequency distribution , mode is the value of x – corresponding to maximum frequency.

x	1	2	3	4	5	6	7	8
f	4	9	16	25	22	15	7	3

For Continuous Frequency distribution:

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

l – lower limit of the modal class

h – magnitude of the modal class

f_1 - frequency of the modal class

$f_0 \; and \; f_2$ - frequencies of the class preceding and succeeding the modal class

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

Hence the maximum frequency is 28.

Thus the modal class is 40-50.

$$\text{Mode} = 40 + \frac{10(28-12)}{2x28-12-20} = 46.67$$

A distribution is having only one mode is called ***Unimodal***.

If it contains more than one mode, it is called ***bimodal*** or ***multimodal***.

Then the value of the mode cannot be determined the above formula and hence mode is *ill-defined*.

$$\boldsymbol{Mode = 3\ Median - 2\ Mean}$$

1) Calculate the mean, median and mode for the following data.

Wages (in lakhs.)	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64
No. of workers	31	47	59	78	104	113	81	60	52	25

Mean = 39.52

Median = 39.77

Mode = 40.6

2) Calculate the Mean, Median and Mode for the following data.

Variable	10-13	13-16	16-19	19-22	22-25	25-28	28-31	31-34	34-37	37-40
Frequency	8	15	27	51	75	54	36	18	9	7

Mean = 24.19

Median = 23.96

Mode = 23.6