Unit III: Summarization Measures:

a. Measures of Central Tendencies:

1. Definition of Average, Types of Averages: Arithmetic Mean, Combined and Weighted mean.
2. Median, and Mode for grouped as well as ungrouped data.
3. Quartiles, Deciles and Percentiles.
4. Using Ogive locate median and Quartiles.
5. Using Histogram locate mode.

b. Measures of Dispersions:

1. Concept and idea of dispersion.
2. Various measures Range, Quartile
3. Deviation: Mean Deviation, Standard Deviation, Variance, Combined Variance.

Miss. Aayesha Ansari
Assistant Professor (Mathematics)

Introduction of Statistics:

- Managerial decision-making can be made efficient and effective by analyzing available data using appropriate statistical tools. Statistical tools not only have application in research (marketing research included) but also in other functional areas like quality management, inventory management, financial analysis, human resource planning and so on.
- **Statistics** originally was meant for collection of facts useful for affaires of the state, like taxes, land records, population demography, etc.
- **Statistics** is the science of data.
- This involves **collecting**, **analyzing** and **interpreting information**.
- "Statistics is a science of Counting". ← Bowley
- "Statistics is the science which deals with collection classification and tabulation of numerical facts as bases for explanation description and comparison of phenomenon." ← Liovelt
- "The science which deals with collection analysis presentation and interpretation of numerical." ← Croxton & Cowden
- Statistical methods are broadly divided into five categories. These are Descriptive Statistics, Analytical Statistics, Inductive Statistics, Inferential Statistics, Applied Statistics.
- **Descriptive Statistics** uses graphical and numerical techniques to summarize and display the information contained in a data set.
- **Inferential Statistics** uses sample data to make decisions or predictions about a larger population of data.

Functions of statistics

1. It simplifies data
2. It facilitates comparison
3. It helps in ascertaining the relation between variables
4. Statistics helps in estimation
5. Statistics helps in Decision making
6. Statistics helps in making decisions where an uncertainty is involved

APPLICATION OF STATISTICS

- Statistical techniques can be used in all fields for which numerical data is collected.
- These are used in social science, like Economic, Psychology, and Sociology as well as in exact science like Biology and Physics.
- Business activity
- Accountancy and Finance
- Social Science
- State Administration
- A common man

LIMITATION OF STATISTICS

- Statistics method can applied to group of data and does not highlight individual facts
- Statistics is a quantitative and Qualitative study of facts
- Statistical conclusions are not true universally
- Statistics can be misused
- Statistics is an indispensable tool of production control and market research.
- Statistical tools are extensively used in business for time and motion study, consumer behaviour study, investment decisions, performance measurements and compensations, credit ratings, inventory management, accounting, quality control, distribution channel design, etc.

We Need Data

- To provide input to survey
- To provide input to study
- To measure performance of service or production process
- To evaluate conformance to standards
- To assist in formulating alternative courses of action
- To satisfy curiosity

It is important to understand the techniques of collection of data because if the data is not collected properly its reliability itself becomes questionable and our entire analysis will be on weak foundations. Based on whether the investigator collects data himself/herself or uses data collected by some other person or agency data is classified into two types. **Primary data and Secondary data**.

SOURCES OF INFORMATION

Primary Source	Secondary Source
• Data is collected by Researcher himself • Data is gathered Through questionnaire, Interviews, observations etc.	• Data collected, Compiled or Written by other Researchers e.g. Books, Journals, newspapers • Any reference must be acknowledge

MEASURES OF CENTRAL TENDANCY

- In many frequency distributions, the tabulated values show a distinct tendency to cluster or to group around a typical central value.
- This behavior of the data to concentrate the values around a central part of distribution is called 'Central Tendency' of the data.
- According to Croxton and Cowden "**An average** is a single value within the range of the data that is used to represent all the values in the series".
- Since an average is somewhere within the range of data, it is sometimes called a **measure of central value.**
- Since all these values in some way or the other represent the central values or average values of the data, they are referred to as **Measures of Central Tendencies.**
- There are three types of measures of central tendencies or Average and they are **mean, median and mode.**

A good measure of central tendency should possess the following characteristics:

- Easy to understand.
- Simple to compute.
- Based on all observations.
- Uniquely defined.
- Possibility of further algebraic treatment.
- Not unduly affected by extreme values.

Arithmetic Mean:

- "Arithmetic mean" or simply "mean" of a given data is defined as the sum of all the values of the data divided by the total number of values.
- The **mean** of n data items $x_1, x_2, \ldots, x_n$, is given by the formula
- $\bar{x} = \frac{\sum_i f_i x_i}{\sum_i f_i}$ **OR** $\bar{x} = \frac{\sum_i x_i}{\sum_i n}$

1. Find the arithmetic mean for the following data representing marks in six subjects 74, 89, 93, 68, 85 and 76.

$$\bar{x} = \frac{\sum_i x_i}{\sum_i n}$$

= (74 + 89 + 93 + 68 + 85 + 76)/6= 485 / 6

$\bar{x}$ **= 80.83**

2. Calculate the mean for the following data:

Age in years:	11	12	13	14	15	16	17
No.ofStudents:	7	10	16	12	8	11	5

Solution:

Age in years (x)	No.of Students (f)	fx
11	7	77
12	10	120
13	16	208
14	12	168
15	8	120
16	11	176
17	5	85
Total	$N = 69$	$\sum fx = 954$

$\sum f = N = 69$, $\sum fx = 954$

$$\bar{x} = \frac{\sum fx}{N} = 13.83 \text{ years.}$$

3. Find the arithmetic mean for the following data representing marks of 60 students.

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No.ofStudents	8	15	13	10	7	4	3

Solution:

Marks	No. of students (f)	Mid-point (x)	fx
10-20	8	15	120
20-30	15	25	375
30-40	13	35	455
40-50	10	45	450
50-60	7	55	385
60-70	4	65	260
70-80	3	75	225
Total	N = 60		$\sum fx = 2270$

$$\bar{x} = \frac{\sum fx}{N} = \frac{2270}{60} = 37.83$$

$\therefore$ The average marks are 37.83

Weighted Mean:

The different weight are assigned to different observations according to their relative importance and then average is calculated by considering weight as well.

This average is called Weighed Arithmetic Mean or Simple weighted mean, denoted by $\overline{x_w}$

$$\overline{x_w} = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

Example: Calculate weighted mean from the following data:

Items	Expenditure (Rs.) (x_i)	Weights (w_i)
Food	290	7.5
Rent	54	2
Clothing	98	1.5
Fuel and Light	75	1
Others	75	0.5

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{\sum wx}{\sum w}$$

Items	Expenditure (Rs.) (x_i)	Weights (w_i)	X_i w_i
Food	290	7.5	2175
Rent	54	2	108
Clothing	98	1.5	147
Fuel and Light	75	1	75
Others	75	0.5	37.5
Total		**12.5**	**2542.5**

$$\bar{x}_w = \frac{\sum wx}{\sum w} = \frac{2542.5}{12.5} = Rs. \quad 203.4$$

Miss. Aayesha Ansari
Assistant Professor (Mathematics)

Combined Arithmetic Mean

For 'k' subgroups of data consisting of 'n_1, n_2, …, n_k' observations (with $\sum_{i=1}^{k} n_i = n$), having respective means, $\bar{x}_1$, $\bar{x}_2$, …, $\bar{x}_k$.
Then combined mean (mean of the all 'k' means) is given by:

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{n}$$

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

Example:

The mean heights and the number of students in three sections of a statistics class are given below:

Sections	Number of students	Mean height (inches)
A	40	62
B	37	58
C	43	61

Calculate overall (or combined) mean height of the students.

Note that we have, n_1=40, n_2=37, n_3=43 and $\bar{x}_1$=62, $\bar{x}_2$=58 and $\bar{x}_3$=61.

So, Combined mean is:
$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = 60.4$ inches

MEDAIN

- Median is another measure of central tendency or another way to express the whole data with the help of a representative number.
- Therefore, median can be defined as the Middle value of a data when the data is arranged in ascending order or descending order.
- The median is not as sensitive to extreme values as the mean

To find the **median** of a group of items:

- Rank the items.
- If the number of items is ***odd***, then the median is the middle item in the list.
- If the number of items is ***even***, the median is the mean of the two middle numbers.

1. Ten students in a math class were polled as to the number of siblings in their individual families and the results were: 3, 2, 2, 1, 1, 6, 3, 3, 4, 2.

   Find the median number of siblings for the ten students.

   Position of the median: 10/2

   Data in order: 1, 1, 2, 2, 2, 3, 3, 3, 4,

   Median=((n+1)/2)th Observation

   = ((9+1)/2)thObservation

   =(10/2)th observation

   = 5th Observation

   Median= 2

- **Median** = $l_1 + \frac{\left(\frac{N}{2} - c.f.\right) * (l_2 - l_1)}{f}$

l_1	:	Lower class limit of Medial class interval
l_2	:	Upper class limit of Medial class interval
f	:	frequency of Medial class interval
c.f.	:	preceding cumulative frequency of less than type of Medial class interval

Miss. Aayesha Ansari
Assistant Professor (Mathematics)

The following data relate to the number of patients visiting a government hospital daily:

No. of patients :	1000-1200	1200-1400	1400-1600	1600-1800	1800-2000	2000-2200
No. of days :	15	21	24	18	12	10

Solution :

No. of patients	**No. of days**	**c.*f*. less than type**
1000-1200	15	15
1200-1400	21	36
1400-1600	24	60
1600-1800	18	78
1800-2000	12	90
2000-2200	10	100
	N=100	

N/2 =100/2=50

See the c.*f* column greater than 50 is 60. The corresponding class interval 1400-1600 is the median class.

$\therefore$ l_1 =1400; l_2=1600; f=24; c.f= 36

$$\therefore M = l_1 + \left[\frac{N/2 - cf}{f}\right](l_2 - l_1) = 1400 + \left[\frac{50-36}{24}\right](1600-1400)$$

$$= 1400+[14]\,(200) = 1400+116.67$$

M **= 1516.67 patients**

Miss. Aayesha Ansari
Assistant Professor (Mathematics)

Mode

- The third measure of central tendency, i.e, **mode**, is again a single value which can represent the whole data.
- Mode is defined to be that number which occurs most frequently in a given data.

$$\text{Mode} = l + \frac{(f_m - f_0)*(l_2 - l_1)}{2f_m - f_0 - f_1}$$

Where, l_1 : lower class limit of Modal class interval

l_2 : Upper class limit of Modal class interval

f_m : Maximum frequency

f_0 : Preceding frequency of modal class interval

f_1 : succeeding frequency of modal class interval

Find the mode for the following data.

Marks:	10-19	20-29	30-39	40-49	50-59	60-69	70-79
No of Students:	8	22	31	44	15	13	10

Solution:

Marks	Class boundaries	No. of Students
10-19	9.5-19.5	8
20-29	19.5-29.5	22
30-39	29.5-39.5	31 f_o
40-49	39.5-49.5	44 f_1
50-59	49.5-59.5	15 f_2
60-69	59.5-69.5	13
70-79	69.5-79.5	10

The maximum frequency is 44.

$\therefore$ The modal class is 39.5-49.5

$l_1 = 39.5$ $f_0 = 31$

$l_2 = 49.5$ $f_1 = 44$, $f_2 = 15$

$$Z = l_1 + \left[\frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)}\right](l_2 - l_1)$$

$$= 39.5 + \left[\frac{44-31}{(44-31)+(44-15)}\right](49.5 - 39.5) = 39.5 + \left[\frac{13}{13+29}\right] \text{x } 10$$

$$= 39.5 + 3.095$$

$$= 42.595$$

$$Z = 42.60$$

Modal Marks = 42.60

EMPIRICAL RELATION BETWEEN MEDAIN AND MODE

For moderately asymmetrical distribution (or for asymmetrical curve), the relation

Mean – Mode = 3 (Mean - Median),

approximately holds. In such a case, first evaluate mean and median and then modeis determined by

Mode = 3 Median – 2 Mean.

If in the asymmetrical curve the area on the left of mode is greater than area on the right then

Mean < median < mode, i. e., (M <Md< M0)

Miss. Aayesha Ansari
Assistant Professor (Mathematics)

MERITS OF A.M.	DEMERITS OF A.M.
1. The mean summarizes all the information in the data	1. It can neither be determined by inspection or by graphical location
2. It is the average of the all observation	2. Arithmetic mean cannot be computed for qualitative data like data on intelligence honesty and smoking habit etc.
3. It is the single point that can be viewed as the point where all the weight of the observation is concentrated	3. It is too much affected by extreme observations and hence it is not adequately represent data consisting of some extreme point
4. It is the center of mass of the data	4. Arithmetic mean cannot be computed when class intervals have open ends

MERITS OF MEDIAN	DEMERITS OF MEDIAN
1) There is a unique median for each data set.	1) Lack of representative character Median fails to be a representative measure in case of such series the different values of which are wide apart from each other. Also, median is of limited representative character as it is not based on all the items in the series.
2) It is not affected by extremely large or small values and is therefore a valuable measure of central tendency when such values occur.	2) Unrealistic:- When the median is located somewhere between the two middle values, it remains only an approximate measure, not a precise value.
3) It can be computed for ratio-level, interval-level, and ordinal-level data. 4) It can be computed for an open-ended frequency distribution if the median does not lie in an open-ended class.	3) Lack of algebraic treatment: - Arithmetic mean is capable of further algebraic treatment, but median is not. For example, multiplying the median with the number of items in the series will not give us the sum total of the values of the series.

Miss. Aayesha Ansari
Assistant Professor (Mathematics)

Merits of Mode:	**Demerits Of Mode:**
(1) Simple and popular: - Mode is very simple measure of central tendency. Sometimes, just at the series is enough to locate the model value. Because of its simplicity, it s a very popular measure of the central tendency. (2) Less effect of marginal values: - Compared top mean, mode is less affected by marginal values in the series. Mode is determined only by the value with highest frequencies. (3) Graphic presentation:- Mode can be located graphically, with the help of histogram. (4) Best representative: - Mode is that value which occurs most frequently in the series. Accordingly, mode is the best representative value of the series. (5) No need of knowing all the items or frequencies: - The calculation of mode does not require knowledge of all the items and frequencies of a distribution. In simple series, it is enough if one knows the items with highest frequencies in the distribution.	Following are the various demerits of mode: (1) Uncertain and vague: - Mode is an uncertain and vague measure of the central tendency. (2) Not capable of algebraic treatment: - Unlike mean, mode is not capable of further algebraic treatment. (3) Difficult: - With frequencies of all items are identical, it is difficult to identify the modal value. (4) Complex procedure of grouping:- Calculation of mode involves cumbersome procedure of grouping the data. If the extent of grouping changes there will be a change in the model value. (5) Ignores extreme marginal frequencies:- It ignores extreme marginal frequencies. To that extent model value is not a representative value of all the items in a series.

First quartile : $Q_1 = l_1 + \frac{\left(\frac{N}{4} - c.f.\right) * (l_2 - l_1)}{f}$

Where,

, l_1 : lower class limit of first quartile class interval

l_2 : Upper class limit of first quartile class interval

c.f. : preceding cumulative frequency of less than type of 1st quartile class interval

f : corresponding frequency of 1st quartile class interval

Third Quartile : $Q_3 = l_1 + \frac{\left(3 * \frac{N}{4} - c.f.\right) * (l_2 - l_1)}{f}$

Where,

l_1 : lower class limit of third quartile class interval

l_2 : Upper class limit of third quartile class interval

c.f. : preceding cumulative frequency of less than type of 3rd quartile class interval

f : corresponding frequency of 3rd quartile class interval

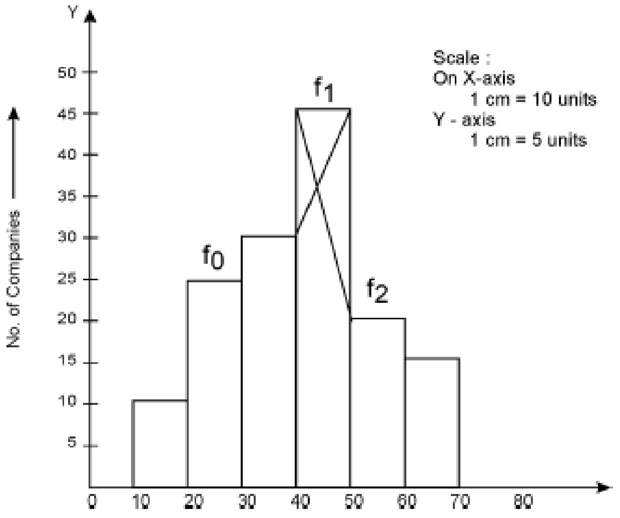
Similarly:

ith Decile : $D_i = l_1 + \frac{\left(i * \frac{N}{10} - c.f.\right) * (l_2 - l_1)}{f}$

ithPercentile : $P_i = l_1 + \frac{\left(i * \frac{N}{100} - c.f.\right) * (l_2 - l_1)}{f}$

Draw a histogram for the following data and hence locate mode graphically.

Profit (in lakhs):	10-20	20-30	30-40	40-50	50-60	60-70
No. of : Companies	12	25	31	45	20	15



Mode Profit (in lakhs)

Modal profits = Rs. 44 lakhs.

Draw a 'less than " cumulative frequency curve for the following data and hence locate the three quartiles graphically.

Age in yrs.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of persons	15	13	25	22	25	10	5	5

Solution :

Age (in yrs.)	No. of persons.	c.*f*. less than type
0-10	15	15
10-20	13	28
20-30	25	53
30-40	22	75
40-50	25	100
50-60	10	110
60-70	5	115
70-80	5	120
	N=120	

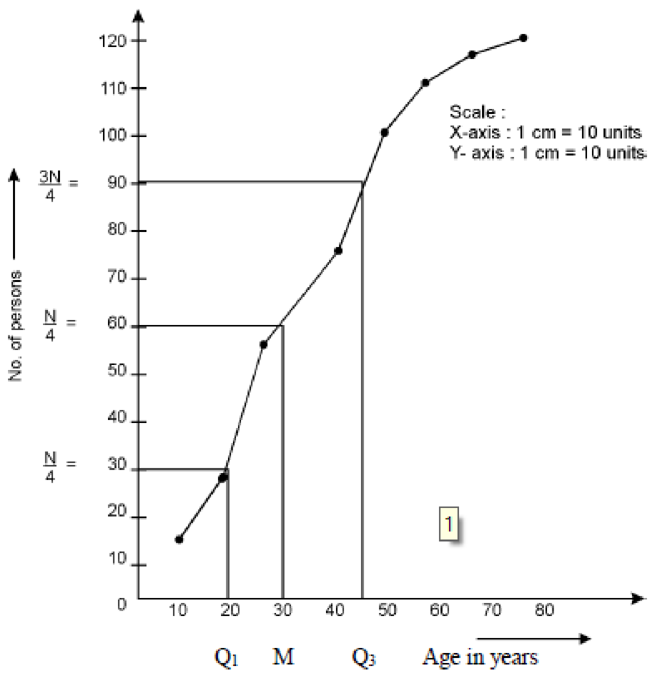
X -axis (upper limit)	Y- axis c.*f*. < type
10	15
20	28
30	53
40	75
50	100
60	110
70	115
80	120

$n/4 = 120/4 = 30$

$n/2 = 120/2 = 60$

$3n/4 = 3\ (120/4)$
$3n/4 = 90$

Less than O-give Curve



From the graph : Q1 = 21 approximately
M = 33 approximately
Q3 = 46 approximately

Miss. Aayesha Ansari
Assistant Professor (Mathematics)