Notes Prepared by *Prof (Mrs) M.J. Gholba*

Measures of Central Tendency

According to Prof Bowley "Measures of central tendency (averages) are statistical constants which enable us to comprehend in a single effort the significance of the whole."

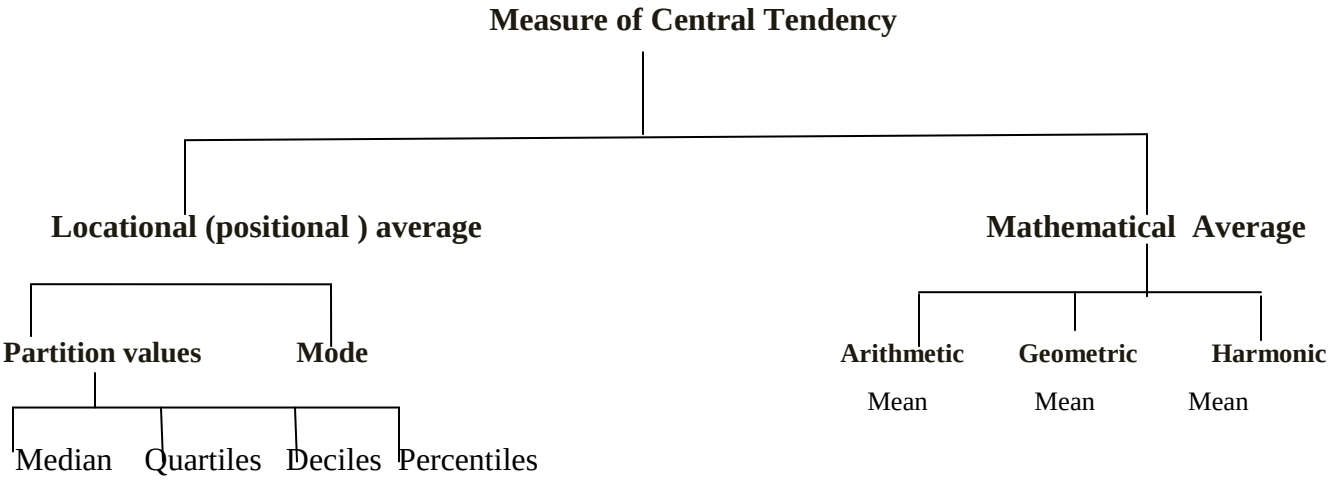
The main objectives of **Measure of Central Tendency are**

1) **To condense data in a single value.**
2) **To facilitate comparisons between data.**

There are different types of averages, each has its own advantages and disadvantages.

Requisites of a Good Measure of Central Tendency:

1. It should be rigidly defined.
2. It should be simple to understand & easy to calculate.
3. It should be based upon all values of given data.
4. It should be capable of further mathematical treatment.
5. It should have sampling stability.
6. It should be not be unduly affected by extreme values.



Partition values: The points which divide the data in to equal parts are called Partition values.

Median: The point or the value which divides the data in to two equal parts., or when the data is arranged in numerical order

The data must be ranked (sorted in ascending order) first. The median is the number in the middle. Depending on the data size we define median as:

It is the middle value when data size N is odd. It is the mean of the middle two values, when data size N is even.

Ungrouped Frequency Distribution

Find the cumulative frequencies for the data. The value of the variable corresponding to which a cumulative frequency is greater than (N+1)/2 for the first time. (Where N is the total number of observations.)

Example 1: Find the median for the following frequency distribution.

X	1	2	3	4	**5**	6	7	8	9
Freq	8	10	11	16	20	25	15	9	6

Solution: Calculate cumulative frequencies less than type.

X	1	2	3	4	**5**	6	7	8	9
Freq	8	10	11	16	20	25	15	9	6
Cum freq	8	18	29	45	**65**	90	105	114	120

N=120, (N+1)/2=60.5 this value is first exceeded by cumulative frequency 65 , this value is corresponding to X-value 5, hence median is 5

Grouped Frequency Distribution First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is greater than (N)/2 for the first time. (N is the total number of observations.) Then that class is median class. Then median is evaluated by interpolation formula.

$$median = l_1 + (l_2 - l_1)\frac{(\frac{N}{2} - cf)}{f_m}$$

Where l_1= lower limit of the median class, l_2= upper limit of the median class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the median class.

f_m= frequency of the median class.

Quartiles : The data can be divided in to four equal parts by three points. These three points are known as quartiles. The quartiles are denoted by Q_i , i = 1,2,3

Qi is the value corresponding to $(iN/4)^{th}$ observation after arranging the data in the increasing order.

For grouped data : First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is greater than (iN)/4 for the first time. (Where N is total number of observations.). Then that class is Qi class. Then Qi is evaluated by interpolation formula.

$$Qi = l_1 + (l_2 - l_1)\frac{(\frac{iN}{4} - cf)}{f_q} \qquad \text{i= 1, 2, 3}$$

Where l_1= lower limit of the Qi class, l_2= upper limit of the Qi class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the Qi class.

f_q= frequency of the Qi class.

Deciles are nine points which divided the data in to ten equal parts.

Di is the value corresponding to $(iN/10)^{th}$ observation after arranging the data in the increasing order.

For grouped data :First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is greater than (iN)/10 for the first time. (Where N is total number of observations.). Then that class is Di class. Then Di is evaluated by interpolation formula.

$$Di = l_1 + (l_2 - l_1)\frac{(\frac{iN}{10} - cf)}{f_d} \qquad \text{i= 1, 2, ……10.}$$

Where l_1= lower limit of the Di class, l_2= upper limit of the Di class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the Di class.

f_d= frequency of the Di class.

Percentiles are ninety-nine points which divided the data in to hundred equal parts.

Pi is the value corresponding to $(iN/100)^{th}$ observation after arranging the data in the increasing order.

For grouped data : First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is greater than (iN)/100 for the first time. (Where N

is total number of observations.) Then that class is Pi class. Then Pi is evaluated by interpolation formula.

$$Pi = l_1 + (l_2 - l_1)\frac{(\frac{iN}{100} - cf)}{f_p}$$

Where l_1= lower limit of the Pi class, l_2= upper limit of the Pi class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the Pi class.

f_p= frequency of the Pi class.

Graphical method for locating partition values: These partition values can be located graphically by using ogives. The point of intersection of both ogives is median.

To locate quartiles, mark N/4 on Y- axis, from that point draw a line parallel to X-axis, it cuts less than type ogive at Q_1 and intersects greater than or equal to curve at Q_3.

To locate Di mark iN/10 on Y-axis , from that point draw line parallel to X-axis, it intersects less than type curve at Di.

Similarly to locate Pi mark iN/100 on Y-axis , from that point draw line parallel to X-axis, it intersects less than type curve at Pi.

Example **2** . Find the median

Daily wages in Rs.	100-200	200-300	300-400	400-500	500-600	600-700
No of workers	4	6	20	10	5	5

Solution : To locate median class we have to calculate cumulative frequencies.

Daily wages in Rs.	100-200	200-300	300-400	400-500	500-600	600-700
No of workers	4	6	20	10	5	5
Cum Freq	4	10	30	40	45	50

N=50 , N/2= 25 so median class is 300-400

$$median = l_1 + (l_2 - l_1)\frac{\left(\frac{N}{2} - cf\right)}{f_m} = 300 + (400 - 300)\frac{(25-10)}{20} = 300 + 100 * \frac{15}{20} = 375$$

Example **3** : Find the median, Q_1, D8,P65 from the following data.

Marks	0-10	10-30	30-50	50-80	80-90	90-100
No of Students	4	12	20	8	4	2

Solution : To locate median class we have to calculate cumulative frequencies.

Marks	0-10	10-30	30-50	50-80	80-90	90-100
No of Students	4	12	20	8	4	2
Cumulative freq	4	16	36	44	48	50

Here N=50 so N/2=25, hence median class is 30-50

$$median = l_1 + (l_2 - l_1)\frac{\left(\frac{N}{2} - cf\right)}{f_m} = 30 + \frac{(50 - 30)(25 - 16)}{20} = 30 + 20 * \frac{9}{20} = 39$$

Here N=50 so N/4=12.5, hence Q1class is 10-30

$$Q1 = l_1 + (l_2 - l_1)\frac{\left(\frac{N}{4} - cf\right)}{f_q} = 10 + \frac{(30 - 10)(12.5 - 4)}{12} = 10 + 20 * \frac{8.5}{12} = 24.16$$

Here N=50 so 8*N/10=40 , hence D8 class is 50-80

$$D8 = l_1 + (l_2 - l_1)\frac{\left(\frac{8N}{10} - cf\right)}{f_{\mathrm{d}}} = 50 + (80 - 50) * \frac{40 - 36}{8} = 50 + 30 * \frac{4}{8} = 65$$

Here N=50 so 65*N/100=32.5 , hence P65 class is 30-50

$$P65 = l_1 + (l_2 - l_1)\frac{(\frac{65N}{100} - cf)}{f_p} = 30 + (50 - 30) * \frac{32.5 - 16}{20} = 30 + 20 * \frac{16.5}{20} = 46.5$$

Use the median to describe the middle of a set of data that *does* have an outlier.

Merits of Median

1. It is rigidly defined.
2. It is easy to understand & easy to calculate.
3. It is not affected by extreme values.
4. Even if extreme values are not known median can be calculated.
5. It can be located just by inspection in many cases.
6. It can be located graphically.
7. It is not much affected by sampling fluctuations.
8. It can be calculated for data based on ordinal scale.

Demerits of Median

1. It is not based upon all values of the given data.
2. For larger data size the arrangement of data in the increasing order is difficult process.
3. It is not capable of further mathematical treatment.
4. It is insensitive to some changes in the data values.

MODE

The mode is the most frequent data value. Mode is the value of the variable which is predominant in the given data series. Thus in case of discrete frequency distribution, mode is the value corresponding to maximum frequency. Sometimes there may be no single mode if no one value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or more than three modes (multi-modal).

For grouped frequency distributions, the modal class is the class with the largest frequency. After identifying modal class mode is evaluated by using interpolated formula. This formula is applicable when classes are of equal width.

$$mode = l_1 + (l_2 - l_1)\frac{d1}{d1 + d2}$$

Where l_1= lower limit of the modal class,

l_2= upper limit of the modal class'

d1 =fm-f_0 and d2=fm-f_1

where fm= frequency of the modal class,

f_0 = frequency of the class preceding to the modal class,

f_1= frequency of the class succeeding to the modal class.

Mode can be located graphically by drawing histogram.

Steps:

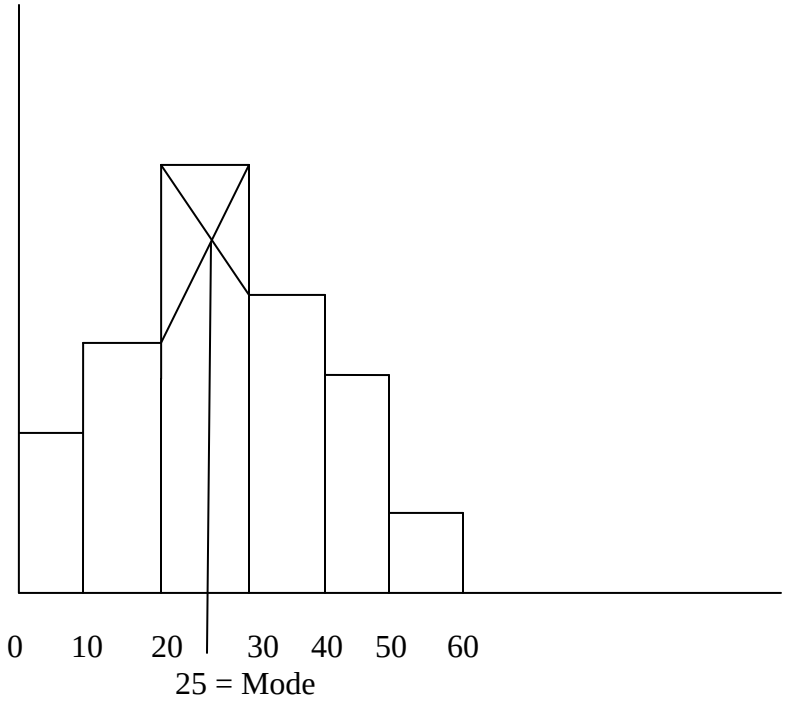
1) Draw histogram
2) Locate modal class (highest bar of the histogram
3) Join diagonally the upper end points of the end points of the highest bar to the adjacent bars.
4) Mark the point of intersection of the diagonals.
5) Draw the perpendicular from this point on the X-axis .
6) The point where the perpendicular meets X-axis gives the modal value.

Example 4: Find the mode

Classes	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	12	18	**27**	20	17	6

Modal class : 20-30 d1= fm-f0= 27-18=9 d2= fm-f1= 27-20=7

$$mode = l_1 + (l_2 - l_1)\frac{d1}{d1 + d2} = 20 + (30 - 20)\frac{9}{9 + 7} = 20 + 10 * \frac{9}{16} = 25.6$$



Use the mode when the data is non-numeric or when asked to choose the most popular item.

Merits of Mode

1. It is easy to understand & easy to calculate.

2. It is not affected by extreme values or sampling fluctuations.
3. Even if extreme values are not known mode can be calculated.
4. It can be located just by inspection in many cases.
5. It is always present within the data.
6. It can be located graphically.
7. It is applicable for both qualitative and quantitative data.

Demerits of Mode

1. It is not rigidly defined.
2. It is not based upon all values of the given data.
3. It is not capable of further mathematical treatment.

Arithmetic Mean

This is what people usually intend when they say "average"

Sample mean: If $X_1, X_2, \ldots\ldots\ldots\ldots\ldots X_n$ are data values then arithmetic mean is given by

$$\bar{X} = \frac{X_1 + X_2 + \cdots . + X_n}{n} = \frac{\sum_1^n Xi}{n}$$

Frequency Distribution: Let $X_1, X_2, \ldots\ldots\ldots\ldots\ldots X_n$ are class marks and the corresponding frequencies are $f_1, f_2, \ldots\ldots\ldots\ldots f_n$, then arithmetic mean is given by

$$N = \sum fi$$

$$\bar{X} = \frac{\sum_1^n fiXi}{N}$$

Example 5 : The Marks obtained in 10 class tests are 25, 20, 20, 9, 16, 10, 21, 12, 8, 13.

The mean $= \bar{X} = \frac{25+20+20+9+16+10+21+12+8+13}{10} = \frac{154}{10} = 15.4$

Example 6 : Find the mean

Xi	9	10	11	12	13	14	15	16
Freq=fi	2	5	12	17	14	6	3	1

Then $N = \sum fi = 60$, and $\sum fi\, Xi = 731$

$$\bar{X} = \frac{\sum_1^n fiXi}{N} = \frac{731}{60} = 12.18$$

Example 7 : The following data represents income distribution of 100 families, Calculate mean income of 100 families.

Income in '00 Rs.	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of families	8	12	25	22	16	11	6

Solution: We have

Income in '00 Rs.	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Class Mark Xi	35	45	55	65	75	85	95
No. of families fi	8	12	25	22	16	11	6

We get N= $\sum$ fi =100 , $\sum$ fi Xi= 6330

Mean =$\bar{X} = \frac{\sum_1^n fiXi}{N} = \frac{6330}{100} = 63.30$

Properties of Mean:

1) Effect of shift of origin and scale.
   If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values . New values U are obtained by shifting the origin to 'a' and changing scale by 'h'
   $Ui = \frac{Xi-a}{h}$ then Mean= $\bar{X} = a + h\bar{U}$
2) Algebraic sum of deviations of set of values taken from their mean is zero.
   a. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values then
      $\sum_1^n (Xi - \bar{X}) = 0$
   b. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values with corresponding frequencies f1, f2,……fn then
      $$\sum_1^n fi(Xi - \bar{X}) = 0$$
3) The sum of squares of deviation of set of values about its mean is minimum.

   $\sum_1^n ((Xi - \bar{X})^2) < \sum_1^n (Xi - A)^2$ where $A \neq \bar{X}$

4) If $Zi = Xi \pm Yi$ i= 1,2 …..n
   $then \;\; \bar{Z} = \bar{X} \pm \bar{Y}$

5) If $\overline{X1}\ and\ \overline{X2}$ are the means of two sets of values containing n_1 and n_2 observations respectively then the mean of the combined data is given by $\bar{X} = \frac{n_1\overline{X1}+n_2\overline{X2}}{n_1+n_2}$
This formula can be extended for k sets of data values as
$$\bar{X} = \frac{n_1\overline{X1} + n_2\overline{X2} + \cdots . + n_k\overline{Xk}}{n_1 + n_2 + \cdots \ldots + n_k}$$

Merits of Mean

1. It is rigidly defined.
2. It is easy to understand & easy to calculate.
3. It is based upon all values of the given data.
4. It is capable of further mathematical treatment.
5. It is not much affected by sampling fluctuations.

Demerits of Mean

1. It cannot be calculated if any observations are missing.
2. It cannot be calculated for the data with open end classes.
3. It is affected by extreme values.
4. It cannot be located graphically.
5. It may be number which is not present in the data.
6. It can be calculated for the data representing qualitative characteristic.

Empirical formula: For symmetric distribution Mean, Median and Mode coincide. If the distribution is moderately asymmetrical the Mean, Median and Mode satisfy the following relationship

Mean- Mode =3(Mean- Median)

Or Mode=3Median-2Mean

Weighted mean : If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values with corresponding weights $W_1, W_2, \ldots\ldots W_n$ then the weighted mean is given by

$$\bar{X}_w = \frac{\sum_1^n WiXi}{\sum_1^n Wi}$$

The mean of a frequency distribution is also the weighted mean.

Use the mean to describe the middle of a set of data that *does not* have an outlier.

Geometric Mean:

a. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values then

$$GM = \sqrt[n]{X1 * X2 \ldots\ldots Xn}$$

Or GM= antilog($\frac{\sum_1^n logXi}{n}$)

b. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values with corresponding frequencies f1, f2,……fn then
if N=∑fi

$$\boldsymbol{GM = \sqrt[N]{X1^{f1} * X2^{f2} * \ldots\ldots Xn^{fn}}}$$

GM= antilog($\frac{\sum_1^n fi\ logXi}{N}$)

Merits of Geometric Mean

1. It is based upon all values of the given data.
2. It is capable of further mathematical treatment.
3. It is not much affected by sampling fluctuations.

Demerits of Geometric Mean

1. It is not easy to understand & not easy to calculate
2. It is not well defined.
3. If anyone data value is zero then GM is zero.
4. It cannot be calculated if any observations are missing.
5. It cannot be calculated for the data with open end classes.
6. It is affected by extreme values.
7. It cannot be located graphically.
8. It may be number which is not present in the data.
9. It cannot be calculated for the data representing qualitative characteristic

Harmonic Mean:

a. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values then Harmonic Mean is given by

$$HM = \frac{n}{\sum \frac{1}{Xi}}$$

b. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values with corresponding frequencies f1, f2,……fn then Harmonic Mean given by
if N=∑fi

$$HM = \frac{N}{\sum \frac{fi}{Xi}}$$

Merits of Harmonic Mean

1. It is rigidly defined.
2. It is easy to understand & easy to calculate.
3. It is based upon all values of the given data.
4. It is capable of further mathematical treatment.
5. It is not much affected by sampling fluctuations.

Demerits of Harmonic Mean

1. It is not easy to understand & not easy to calculate.
2. It cannot be calculated if any observations are missing.
3. It cannot be calculated for the data with open end classes.
4. It is usually not a good representative of the data.
5. It is affected by extreme values.
6. It cannot be located graphically.
7. It may be number which is not present in the data.
8. It can be calculated for the data representing qualitative characteristic.

Selection of an average:

No single average can be regarded as the best or most suitable under all circumstances. Each average has its merits and demerits and its own particular field of importance and utility. A proper selection of an average depends on the 1) nature of the data and 2) purpose of enquiry or requirement of the data.

A.M. satisfies almost all the requisites of a good average and hence can be regarded as the best average but it cannot be used

1) in case of highly skewed data.
2) in case of uneven or irregular spread of the data.
3) in open end distributions.
4) When average growth or average speed is required.
5) When there are extreme values in the data.

Except in these cases AM is widely used in practice.

Median: is the best average in open end distributions or in distributions which give highly skew or j or reverse j type frequency curves. In such cases A.M. gives unnecessarily high or low value whereas median gives a more representative value. But in case of fairly symmetric distribution there is nothing to choose between mean, median and mode, as they are very close to each other.

Mode : is especially useful to describe qualitative data. According to Freunel and Williams, consumer preferences for different kinds of products can be compared using modal preferences as we cannot compute mean or median. Mode can best describe the average size of shoes or shirts.

G.M. is useful to average relative changes, averaging ratios and percentages. It is theoretically the best average for construction of index number. But it should not be used for measuring absolute changes.

H.M. is useful in problems where values of a variable are compared with a constant quantity of another variable like time, distance travelled within a given time, quantities purchased or sold over a unit.

In general we can say that A.M. is the best of all averages and other averages may be used under special circumstances.