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 observation shows a tendency to cluster around a single value. This property is called central tendency and the value around which the data cluster is called Average or measure of central tendency

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

1) To condense data in a single value.

2) To facilitate comparisons between data.

The commonly used measures of central tendency are Mean, Mode and Median. That is 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.

Arithmetic Mean

There are two type of Arithmetic Mean, simple arithmetic mean and weighted arithmetic mean

Simple Arithmetic Mean
Raw Data

If $x_1, x_2, \ldots, x_n$ are data values then arithmetic mean is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ungrouped Frequency Distribution

If $x_1, x_2, \ldots, x_n$ are data values and $f_1, f_2, \ldots, f_n$ are the frequencies then arithmetic mean is given by

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_3}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i\, f_i$$

where N= $\sum_{i=1}^{n} f_i$.

Frequency Distribution

If $x_1, x_2, \ldots, x_n$ are class marks or mid values and $f_1, f_2, \ldots, f_n$ are the frequencies then arithmetic mean is given by

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_3}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i$$

Effect of shift of origin and scale on mean

If $x_1, x_2, \ldots, x_n$ are data values and a and c (>0) two real numbers. Define $u_i = \frac{x_i - a}{c}$ and let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$. Then $\bar{x} = c\bar{u} + a$.

Combined Mean

If $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ are the means of k sets of values containing $n_1, n_2, \ldots, n_k$ observations respectively then the mean of the combined data is given by

$$\bar{X} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k}$$

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.

Weighted mean :

If $x_1, x_2, \ldots, x_n$ are the data values and with corresponding weights $w_1, w_2, \ldots, w_n$ then the weighted mean is given by

$$\bar{X} = \frac{\sum w_i x_i}{\sum w_i}$$

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. That is the median is middle most value which divide the entire data in to two equal parts.

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. That is it is the (n+1)/2 th observation when the data is arranged in order. 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 is the median. Where N is the total number of observations.

For grouped Frequency Distribution

First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is N/4 this class is the median class. Then Median is given by

$$M_e = l + \frac{c\left(\frac{N}{4} - m\right)}{f}$$

Where l = lower boundary of the median class, c = class interval of the median class
N= Number of observations, m = cumulative frequency of the class proceeding to the median class. f = frequency of the median class.

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. 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 the raw data mode is that value which repeat most number of times. Thus if more than one observation repeated same number of times and is the most then we have more than one mode. If no observation repeated more than one time then the mode is ill-defined.

For the discrete frequency distribution, mode is the value corresponding to maximum frequency.

For grouped frequency distributions, the modal class is the class with the largest frequency. After identifying modal class mode is evaluated by using formula.

$$M_0 = l + \frac{c(f_1 - f_0)}{2f_1 - f_2 - f_0}$$

Where l = lower boundary of the modal class,
f_1= frequency of the modal class,
f_0= frequency of the class preceding to the modal class,
f_2= frequency of the class succeeding to the modal class.
c = class interval of the modal class.

The following formula can be used for calculation of the mode

$$M_0 = l + \frac{cf_2}{f_1 + f_2}$$

Mode can also 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.

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.

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

Geometric Mean:

Raw Data

If $x_1, x_2, \ldots, x_n$ are data values then geometric mean is given by

$$GM = \sqrt[n]{x_1 x_2 \ldots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

Working formula GM = $\boldsymbol{antilog\left(\frac{1}{n}\sum \log x_i\right)}$

Ungrouped Frequency Distribution

If $x_1, x_2, \ldots, x_n$ are data values and $f_1, f_2, \ldots, f_n$ are the frequencies then geometric mean is given by

$$GM = \left(\prod_{i=1}^{n} {x_i}^{f_i}\right)^{\frac{1}{N}}$$

where N= $\sum_{i=1}^{n} f_i$.

Frequency Distribution

If $x_1, x_2, \ldots, x_n$ are class marks or mid values and $f_1, f_2, \ldots, f_n$ are the frequencies then geometric mean is given by

$$GM = \left(\prod_{i=1}^{n} {x_i}^{f_i}\right)^{\frac{1}{N}}$$

where N= $\sum_{i=1}^{n} f_i$.

Working formula GM = $\boldsymbol{antilog\left(\frac{1}{N}\sum f_i log\ x_i\right)}$

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:

The simple harmonic mean of a set of observation is defined as the reciprocal of the A.M. of the reciprocals of the observations.

Raw Data

If $x_1, x_2, \ldots, x_n$ are data values then harmonic mean is given by

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \sum_{i=1}^{n} \frac{n}{\frac{1}{x_i}}$$

Ungrouped Frequency Distribution

If $x_1, x_2, \ldots, x_n$ are data values and $f_1, f_2, \ldots, f_n$ are the frequencies then harmonic mean is given by

$$HM = \frac{N}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_n}{x_n}} = \frac{N}{\sum \frac{f_i}{x_i}}$$

where N= $\sum_{i=1}^{n} f_i$.

Frequency Distribution

If $x_1, x_2, \ldots, x_n$ are class marks or mid values and $f_1, f_2, \ldots, f_n$ are the frequencies then harmonic mean is given by

$$HM = \frac{N}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_n}{x_n}} = \frac{N}{\sum \frac{f_i}{x_i}}$$

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.