• Introduction and Various Measures of Dispersion

Welcome to the series of eLearning module on Measures of Dispersion. In the previous modules, we have seen the basic concept of dispersion. Here we will discuss about the various measures of dispersion, their merits and demerits.

The main theme of this module is dispersion or variability, which provides us one more step in increasing our understanding of the pattern of data.

At the end of this session, you will be able to understand and explain:

- Explain what is dispersion and purpose of measuring dispersion
- Explain what are the different measures of dispersion and types of measures of dispersion
- Explain how to calculate the measures of dispersion
- Explain merits and demerits of dispersion

Let us do a quick recap of what does dispersion mean.
The measure of central tendency are used to estimate "normal" values of a dataset, whereas the measures of dispersion are important for describing the spread of the data, or its variations around a central value.

The measure of dispersion appears to serve two purposes.

- It is one of the most important quantities used to characterize a frequency distribution.
- It affords a basis of comparison between two or more frequency distributions.

If we know the average alone, we cannot form a complete idea about the distribution. This will be clear to you from the following example:

Let us take an example of a group of students, which are divided into three groups. Group X, Group Y and Group Z.

Figure 1

Students	Group X	Group Y	Group Z
1	7	3	1
2	8	6	5
3	**9**	**9**	**9**
4	10	12	13
5	11	15	17
N = 5	**45**	**45**	**45**

In all these cases we see that N the number of students is 5 and the mean is 9. If we are given that the mean of 5 observations is 9, we cannot form an idea as to whether it is the average of the first ,second or the third series or of any other series of 5 observations whose sum is 45. Thus we see that the measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures. One such measure is dispersion.

Now we will discuss about the various measures of dispersion. Measures of dispersion is classified into two broad categories:

- Absolute Measures of Dispersion: The absolute measures of dispersion can be compared with one another only if the two belong to the same population and are expressed in the same units like inches, rupees etc.
- Relative Measures of Dispersion: The relative measures of dispersion can be found only by calculating:
  Positional Measure: Coefficient of range and coefficient of quartile deviation.
  Calculated Measures: Coefficient of Mean Deviation and Coefficient of Standard Deviation.

In other words:

- The first measures of dispersion express the spread of observations in terms of distance between the values of selected observations. These are also termed as distance measures, example- range and inter quartile range (or quartile deviation).
- The second measures of dispersion express the spread of observation in terms of the average of deviation of observation from some central value,

example- mean deviation and standard deviation.

This hierarchy helps you better to understand how all the four types of measures of dispersion are divided into categories.

Here, the absolute measure of dispersion is divided into two parts- Based on selected observations and based on all the observations.

Range and Inter quartile range which is also known as quartile deviation comes under based on selected observations.

Whereas, Mean deviation and standard deviation comes under based on all the observations.

2. Range

Range: The simplest measure of dispersion is the range, which is the difference between the maximum value and the minimum value of data. We can use range when the observation is less than five and data are on the ordinal scale.

For example: If A is the greatest and B is the smallest observation in a distribution, then range is given by:

Range is $X_{max} – X_{min}$

That is, Range is A-B.

Let us take another example of a height of football players.
Here, in the first team of five players the minimum height is 72 and the maximum height is 78. Thus, the range is 6.

Similarly, in the second team the minimum height of a player is 67 and maximum is 84. Thus, the range is 17.

In order to explain these examples in detail we have taken one more example of set of data, i.e. 12, 10, 5, 15, 11, and 20.

As we know that range is given by: $X_{max} – X_{min}$

Here, in this given set of data, X_{max} is 20, and X_{min} is 5.

Thus, Range is 20-5 = 15.

Merits of range are:

- It is the simplest but a crude measure of dispersion.
- It is easy to calculate
- It is mainly used in situations where one wants to quickly have some idea of the variability of a set of data.
- When the sample size is very small, the range is considered quite adequate as a measure of the variability. Thus, it is widely used in quality control where a continuous check is needed
- The range is also a suitable measure in weather forecast

There are some major demerits of range, they are:

- It is based only on two items and does not cover all the items in a distribution
- It is subject to wide fluctuation from sample to sample based on the same population.
- It also fails to give any idea about the pattern of distribution
- In case of open-ended distributions, it is not possible to compute the range

Inter quartile range: The inter quartile range denotes the difference between the third and the first quartiles.

Symbolically, inter quartile range is $Q_3 - Q_1$

The semi-inter quartile range (SIR) or quartile deviation is defined as the difference of the third and the first quartiles divided by two.

It is a better measure than the range as it makes use of 50% of data. But since it ignores other 50% of the data, it cannot be regarded as a reliable measure.

Quartile divides a series into four equal parts. One fourth (25%) of total items fall below. Three fourth (75%) of total items fall above.

Here, the first quartile is the 25^{th} Percentile and third quartile is the 75^{th} percentile. Thus, Semi inter quartile is Q3-Q1/2.

Let us take one example where the scores of a cricket game in a series are given. Here, calculate what the SIR for the given data is.

Figure 2

Scores	2	4	6	8	10	12	14	20	30	60

In this table 25 % of the scores are below 5. 5 is the first quartile. Whereas 75 % of the scores are above 25. 25 is the third quartile.

Thus, (Q3-Q1)/2 = (25-5)/2 = 10.

Let us take one more example to calculate inter quartile deviation.
Here, data of age in years and no. of teachers is given.

Let us find out the cumulative frequency.
Age in years is (x), no. of teachers is (f).
Cumulative frequency is calculated by multiplying x and f.

Figure 3

Age in Years	No. of Teachers
50	10
51	12
52	15
53	10
54	14
55	18
56	6

Here, Q1 = Size of the (N+1)/4th item.
Q1 = Size of the (85+1)/4 th item.
Thus, 21.5^{th} item is 51.
Similarly, Q3 = Size of the 3(N+1)/4th item.
Q3 = Size of the 3(85+1)/4 th item.
Thus, 64.5^{th} item is 55.

Merits of inter quartile range are:

- As compared to range, it is considered a superior measures of dispersion
- In case of open-end distribution, it is suitable
- It is not influenced by the extreme values in a distribution; it is particularly suitable in highly skewed or erratic distributions.

Demerits of inter quartile range are:

- Like the range, it also fails to cover all the items in a distribution
- It is not suitable for further mathematical treatment
- It varies widely from ample to sample based on the same population

• Mean Deviation

Mean Deviation: The mean deviation is an average of absolute deviations of individual observations from the central value of a series.

Average deviation about mean is: $MD = \frac{\sum \left| x_i - \bar{x} \right|}{N}$ **(Formula 1)**

Where, MD = mean deviation

$\left| x_i - \bar{x} \right|$ = deviation of an item from the mean, ignoring positive and negative signs **(Formula 2)**

N = the total number of observations.

In this example, 72, 85, 87, 89, 90 and 93 are the given set of data.

Here, N = 6 (i.e. number of data).

x = 516 (total of all the data's).

Thus, Mean is 516/6 = 86

Now we will see how to calculate the mean deviation using the formula.

As we know, xi is 516 i.e. the total of the set of data.

Mean is 86.

We have to calculate xi-x bar and |xi-x bar| ignoring all negatives.

Take the first row to explain how to calculate xi-x and |xi-x bar|.

In the first row of the table, x bar is 86. In order to calculate xi-x, we have to subtract 93 – 86. We will get 7 as a result.

Similarly all the rows are calculated.

|xi-x bar| ignoring negative is a deviation of an item from the mean, ignoring

positive and negative signs.

Thus, the total of xi-x bar is 0 and |xi-x bar| is 30.

$$MD = \frac{\sum \left| x_i - \bar{x} \right|}{N}$$ (**Formula 3)**

As per the formula:

|xi-x bar| is 30 and N is 6.
Thus, mean deviation is 30/6 = 5.

Merits of mean deviation are:

- A major advantage of mean deviation is that, it is simple to understand and easy to calculate
- It takes into consideration each and every items in a distribution. As a result, a change in the value of any item will have its effect on magnitude of mean deviation
- The values of extreme items have less effect on the value of the mean deviation

Demerits of mean deviation are:

- It is not capable of further algebraic treatment
- At time it may fails to give accurate results
- Strictly on mathematical considerations, the method is wrong as it ignore the algebraic signs when the deviation are taken from the mean

• Standard Deviation

Standard Deviation: Standard deviation is the positive square root of the arithmetic mean of the squares of the deviation given from their arithmetic mean. Standard deviation, usually denoted by the Greek letter small sigma (σ).

Formula of standard deviation for ungrouped data are:

$$\sigma = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{N}}$$ (**Formula 4)**

Let us take an example: In this table: 20, 15, 19, 24, 16, and 14 are the given data.

Figure 4

Numbers (N)	Data
1	20
2	15
3	19
4	24
5	16
6	14
N = 6	**= 108**

N is 6 which is the numbers of data.
Total of the given data is 108.
Thus, mean is 108/6 = 18

We have to calculate xi-x bar and (xi-x bar) square.
Take the fourth row to explain how to calculate xi-x bar and (xi-x bar)square.
Here, x bar is 18.
xi of fourth row is 24. As per the formula xi-x bar, we will get 24-18 i.e. 6.
Similarly all the rows are calculated.
The result of xi-x bar is then squared to get (xi-x bar)square.
i.e. 6x6 is 36.
Similarly all the rows are calculated.
Thus, the total of (xi-x bar)square is 70.

As per the formula:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$ (**Formula 5)**

Now we will see how to calculate standard deviation for grouped data.

Formula is: $S = C\sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}}$ (**Formula 6)**

Simplified formula or computational formula of standard deviation is:

$$S = C\sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

(Formula 7)

We will take an example of marks of no. of students.

Distribution of marks is given, like 0-10, 10-20, 20-30 and so on.
No. of students is also given, like 1, 3, 6, and so on.

Figure 5

Marks	No. of Students
0-10	1
10-20	3
20-30	6
30-40	10
40-50	12
50-60	11
60-70	6
70-80	3
80-90	2
90-100	1

Here, we have to find out the mid-point, x, fx and (fx)square.

Let us take the last row for the calculation.

Mid point is calculated by adding the range of marks 90+100/2, we will get 95 as mid-point.

X is calculated by subtracting the midpoint with total frequency divided by 10, i.e. 95-55/10 = 4.

Fx is calculated by multiplying x with f. In case of last row fx is 4x1= 4.
On doing the square of 4 we will get (fx)square i.e. 16.
Thus, total fx = -45,
Total (fx)square is 231, and total frequency is 55.

As per the formula:

$$S = C\sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$ (**Formula 8)**

Thus, we see that standard deviation satisfies almost all the properties laid down for an ideal measure of dispersion expect for general nature of extraction the square root which is not readily comprehensible for non-mathematical person. It may also be pointed out that standard deviation gives greater weight to extreme values and such has not found favor with economists or businessmen who are not interested in the results of the modal class. Taking into consideration the pros and cons and also the wide applications of standard deviation in statistical theory, we may regard standard deviation as the best and most powerful measure of dispersion.

Merits of standard deviation are:

- It is based on each and every items of the data
- It take care of both positive and negative deviation
- It is less affected by variations in sampling
- It is most popularly used in statistics

Demerits of standard deviation:

- It is difficult to understand
- It is affected by the extreme values

• Variance

Variance: The square of standard deviation is called as variance.

$$\sigma^2 = \frac{\sum\left(xi - \bar{x}\right)^2}{N}$$ (**Formula 9)**

Formula is :

- The larger the variance is, more the score deviate on average, away from the mean
- The smaller the variance is, less the score deviate on average, away from the mean

When calculating variance, it is often easier to use a computational formula which is algebraically equivalent to the definitional formula :

$$\sigma^2 \frac{1}{n_1 + n_2}\left[n_1\left(\sigma_1^{\ 2} + d_1^{\ 2}\right) + n_2\left(\sigma_2^{\ 2} + d_2^{\ 2}\right)\right]$$ **(Formula 10)**

Let us take an example where the value of x is given like, 20, 15, 19, 24, 16, and 14.
This table has two columns x and x square.
X square is the square of x.
Total of x is 108, total of x square is 2014, and N is 6.

If n_1, n_2 are the sizes, $_1$, $_2$ the means, and σ_1, σ_2 the standard deviation of two series, then the standard deviation σ of the combined series of size n_1+n_2 is given by

$$d_1 = \bar{x}_1 - \bar{x}, d_2 = \bar{x}_2 - \bar{x}$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$ (**Formula 11)**

Whenever we want to compare the variability of the two series which differ widely in their averages or which are measures in different units, we do not merely calculate the measures of dispersion but we calculate the coefficients of dispersion which are pure numbers independent of the units of measurement. The coefficients of dispersion (C.D) based on different measures of dispersion are as follows:

This hierarchy helps you better to understand the types of coefficient of dispersion.
Here, the relative measure of dispersion is divided into two parts- Based on selected observations and based on all the observations.
Coefficient of Range and Coefficient of quartile deviation comes under based on selected observations.
Whereas, Coefficient of Mean deviation and Coefficient of standard deviation comes under based on all the observation.

Coefficient of dispersion based on range:
Coefficient of dispersion based on quartile deviation:
Coefficient of dispersion based on mean deviation:
Coefficient of dispersion based on standard deviation:
Coefficient of variation is the ratio of the standard deviation and the mean expressed in terms of percentage.

In other words, 100 times the coefficient of dispersion based upon the standard deviation is called coefficient of variation (C.V).
According to professor Karl Pearson who suggested this measure, C.V is the percentage variation in the mean, standard deviation being considered as the total variation in the mean.

Here's a summary of our learning in this session:

- What is dispersion and purpose of measuring dispersion
- Different measures of dispersion and types of measures of dispersion
- How to calculate the measures of dispersion
- Merits and demerits of dispersion