**Measures of Dispersion**

Suppose two students, A and B, in a school received in eight monthly examinations the following marks in a particular subject.

|  |  |  |
| --- | --- | --- |
| **Sl.**  **No.** | **Marks obtained by A** | **Marks obtained by B** |
| 1 | 63 | 61 |
| 2 | 47 | 54 |
| 3 | 56 | 56 |
| 4 | 44 | 57 |
| 5 | 66 | 60 |
| 6 | 65 | 59 |
| 7 | 80 | 55 |
| 8 | 43 | 62 |
| **Mean** | **58** | **58** |

The mean score of each student is the same. Yet the overall nature of the scores was not at all the same for the two students.

Student A received as a high score as 80 and as low a score as 43. On the other hand, student B’s scores remained near about 58 throughout. The marks of student B are closed to one another while those of student A are widely dissimilar. This implies that B gave a more consistent performance than A. Thus averages alone are not sufficient to reveal all the characteristics of data.

Thus, in order to give a proper idea about the overall nature of the given values of a variable, it is necessary, besides mentioning the average value, to state how scattered or dispersed the given values are about the average. This feature of a variable is called dispersion.

It may be said that while the central tendency of a variable is the tendency of its values to be *similar*, its dispersion represents the tendency of the values to be *different*.

A measure of dispersion is designed to state numerically the extent to which individual observations vary on the average.

**Desirable properties for a good measure of dispersion**

A good measure of dispersion should obey the following conditions:

* It should be rigidly defined
* It should be based on all observations
* It should be readily comprehendible
* It should be fairly easily calculated
* It should be affected as little as possible by sampling fluctuations
* It should readily lend itself to algebraic treatment

**Different measures of dispersion**

There are several measures of dispersion. These are:

(a) Range

(b) Mean absolute deviation (MAD)

(c) Interquartile range or Quartile deviation

(d) Standard deviation

**Range**

Range of a set of observations is the difference between the maximum and the minimum values, i.e.

Range is the simplest of all measures of dispersion. It is easy to understand. Range is also simple to calculate. However, range has many limitations.

**Limitations of range**

* It does not depend on all observations. The values of intermediate observations are not at all necessary for its calculation.
* It is highly affected by extreme values. Presence of an outlier can increase the range many times.
* It does not take into account the form of the distribution. The same value of range might be obtained from a Bell-shaped or a J-shaped distribution although the nature of dispersion in the two is completely different.
* Range cannot be calculated from frequency distributions with open-end classes.

**Mean absolute deviation (MAD)**

It is the arithmetic mean of absolute deviations (or differences) from an average A (arithmetic mean or median).It signifies average distance between each data value and the average of the data set.

* **Mean absolute deviation for ungrouped data or individual observations**

If are values of a variable , then the mean deviation (MD) from an average A (AM or median) is given by

* **Mean absolute deviation of a discrete frequency distribution**

If the observations have frequencies respectively, then mean deviation from an average A (AM or median) is given by

, where

* **Mean absolute deviation of a grouped or continuous distribution**

If and are the mid-value and frequency of the class then mean deviation from an average A (AM or median) is given by

, where

*It may be noted that the mean deviation is least when measured from the median.* Therefore, it would seem proper to use the median as origin in computing the mean deviation. In practice, however, the mean deviation is generally computed about the arithmetic mean.

**Advantages and Limitations of mean absolute deviation (MAD)**

* Depends on all observations
* Less sensitive to extreme outliers compared to standard deviation
* Procedure of neglecting the signs and taking absolute deviations makes algebraic treatment difficult

**Quartile deviation (or Semi-interquartile range)**

Quantiles are such values (of the variable) which divide the total number of observations into 4 equal parts. Obviously, there are 3 quartiles-

1. First quartiles (or Lower quartile):
2. Second quartile (or Middle quartile):
3. Third quartile (or Upper quartile:

Then, interquartile range (IQR) and quartile deviation are defined as follows:

In the presence of outliers, IQR or quartile deviation is a better representation of the amount of spread in the data rather than the range. This is because in computation of IQR the bottom 25% of the data points and the top 25% of the data points are ignored and thus IQR statistic is more robust with respect to outliers.

Computation of the standard deviation is extremely difficult or impossible when the observations are given in a frequency table with class-intervals of varying width or with one or both of the terminal classes undefined. Then IQR or quartile deviation may be used to represent the measure of dispersion.

**Standard deviation**

Standard deviation is the positive square-root of the arithmetic mean of squares of deviations from arithmetic mean and it is denoted by or . The square of standard deviation is called as variance.

* **Standard deviation of individual observations**

If are values of a variable , then the standard deviation (SD) of is given by

Now,

, since

Hence,

***For the purpose of computation of SD, the above form is used frequently.***

* **Standard deviation of a discrete frequency distribution**

If the observations have frequencies respectively, then standard deviation (SD) of is given by

, where

* **Standard deviation of a grouped or continuous distribution**

If and are the mid-value and frequency of the class, then standard deviation (SD) of is given by

, where

**Some properties of standard deviation**

1. Standard deviation is independent of the change of origin.

***Proof:*** Let be a set of observations and let be the quantities derived from them by shifting the origin to an arbitrary constant , i.e. (). We have to prove that .

By definition, , where

Now, so that

Substituting this,

or,

1. Standard deviation is independent of change of origin, but is dependent on the change of scale.

***Proof:*** Let be a set of observations. If we change the origin of to and the scale to (positive), i.e. write

, then

Hence,

and

or,

or,

It is observed from the result that on the right hand side, the new origin is absent, but the scale is present. This proves that SD is unaffected by any change of origin but depends on scale.

1. Let there be two sets of values of with and values, and let , be their means and and their SDs. Then the SD of for the two sets pooled together ( can be expressed in terms of , , , and , as follows:

***Proof:*** The grand mean of is . Denoting by (=1, 2,...,) and (=1, 2,...,) the values in the two sets, we may write the sum of squares of the deviations of the values from as

where is the total variance of . But

Similarly,

Thus,

Generally, if there be sets with , ,..., , having means , , ..., and , ,..., , the standard deviation of for all the sets taken together is given by

where is the grand mean of .

1. The standard deviation is the minimum root-mean-square deviation.

***Proof:*** Let be a given set of observations. By definition,

(i)

If instead of taking deviations from mean, we consider deviations from an arbitrary constant , then

Root-mean-square deviations from . (ii)

We have to show that (ii) cannot be smaller than (i) whatever be the value of .

Now,

The second term in the above expression can never be negative and therefore,

or,

**Example:** Calculation of SD from group frequency distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Height  (inch) | Frequency | Mid-value  () |  |  |  |
| 60-62 | 35 | 61 | -2 | -70 | 140 |
| 62-64 | 27 | 63 | -1 | -27 | 27 |
| 64-66 | 20 | 65 | 0 | 0 | 0 |
| 66-68 | 13 | 67 | 1 | 13 | 13 |
| 68-70 | 5 | 69 | 2 | 10 | 20 |
| Total | 100 | - | - | -74 | 200 |

If , then

Therefore, , and inches

**Advantages of standard deviation**

* It is rigidly defined.
* It is based on all observations
* It is least affected by fluctuations of sampling
* It is amenable to algebraic treatment

**Disadvantage of standard deviation**

It cannot be used for comparing the dispersion of two or more series of observations given in different units.

**Exercises**

**Ex 1: C**alculate the quartile deviation from the following data:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class interval | 10-15 | 15-20 | 20-25 | 25-30 | 30-40 | 40-50 | 50-60 | 60-70 | Total |
| Frequency | 4 | 12 | 16 | 22 | 10 | 8 | 6 | 4 | 82 |

**Ex 2:** Calculate the appropriate measure of dispersion from the following data:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Wage per week (Rs.) | Less than 35 | 35-37 | 38-40 | 41-43 | Over 43 |
| No. of wage earners | 14 | 62 | 99 | 18 | 7 |

**Ex 3:** Measurements of the lengths in feet of 50 iron rods are distributed as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Class boundary | 2.35-2.45 | 2.45-2.55 | 2.55-2.65 | 2.65-2.75 | 2.75-2.85 | 2.85-2.95 | 2.95-3.05 |
| Frequency | 1 | 4 | 7 | 15 | 11 | 10 | 2 |

Find the value of the mean deviation.

**Ex 4:** Find the mean and the standard deviation of the first natural numbers.

**Ex 5:** The mean and standard deviation (SD), calculated from 20 observations are 15 and 10 respectively. If an additional observation 5, left out through oversight, be included in the calculations, find the corrected mean and SD.

**Ex 6**: For a distribution of 280 observations, mean = 54 and standard deviation = 3. On checking it was discovered that observations which should correctly read as 62 and 82 had been wrongly recorded as 64 and 80 respectively. Calculate the correct values of mean and standard deviation.

**Ex 7:** Find the standard deviation from the following frequency distribution:

|  |  |
| --- | --- |
| **Height in inches** | **No. of students** |
| Over 60 but not more than 62 | 35 |
| Over 62 but not more than 64 | 27 |
| Over 64 but not more than 66 | 20 |
| Over 66 but not more than 68 | 13 |
| Over 68 but not more than 70 | 5 |

**Ex 8:** Prove that the standard deviation calculated from two values and of a variable is equal to half their difference.

**Ex 9:** If , and be the number of observations, standard deviation and mean of a set of observations, and , and be those for a second set of observations, prove that the standard deviation of the combined set of (+ ) observations is given by

(+ )

where , and

**Ex 10:** Show that if is the arithmetic mean of the values (1, 2,..., , then

, where

[**Ans.** E1: 8; E2: 1.75; E3: 0.11; E4: Mean =, SD =; E5: Mean =14.52, SD = 9.99; E6: Mean = 54, SD = 3.04; E7: 2.42 inches]

**Relative measures of dispersion**

All the measures of dispersion are expressed in the same units as those of the variable. As such they cannot be used in comparing two distributions of different types with respect to their variability. For example, a difficulty is encountered when we want to compare the dispersion of a set of height (in cm) with the dispersion of a set of weight (in kg). For purpose of such comparison, therefore, a measure of dispersion has to be made free from the units of the variable. The simplest procedure to obtain such a measure of dispersion is to express a measure of dispersion as a percentage of a measure of central tendency. Such a measure is called as the relative measure of variability. There are three such measures-

Among these, ***coefficient of variation is the most important and is used in almost all cases***.

The relative measure of dispersion may be useful even when we want to compare sets of data expressed in the same units. For example, suppose repeated measurements are being taken of two rods, one of length 10 cm and another of length 100 cm. Let the standard deviation of each set of measurements be 2 cm. But a standard deviation of 2 cm in the first case does not mean the same thing as a standard deviation of 2 cm in the second. For the first set of measurements is then much less accurate than the second set of measurements. In this case, coefficient of variation will give a true picture of their relative accuracy.