

A Note on Karl Pearson's Coefficient of Dispersion

R. Sharma*
R.G. Shandil*
G. Kapoor*

Statistics is a tool of paramount importance in the derivation of meaningful and useful conclusions from a given data. There is a large number of subjects like Economics, Business Management and Biology etc in which quite often we need to analyze a large body of data. In such situations the statistical methods are used to visualize the data by summarizing the information contained in that data. The measure of central tendency and dispersion are two most important features in this context. Three types of averages namely arithmetic mean, median and mode are generally used as the measure of central tendency. An average gives a single value as the representative of the whole set of the data and shows the tendency of the values in the data to be similar. The values in the data may lie very near to average or be widely scattered about the averages. The measure of dispersion gives an idea about the scattering of the values in the data about the average. The range, mean deviation and standard deviation are three important measure of dispersion.

Let A and B represent two sets of data. Suppose we wish to compare the relative variability of the values in A and B . The data given in A and B may have different units of measurement and their averages may differ widely. In this situation the measure of dispersion does not serve the purpose and we have to calculate the coefficient of dispersion. We know that the coefficient of dispersion is a pure number independent of the units of the measurement. A number of coefficients of the dispersion are studied in the literature. The Karl Pearson Coefficient of dispersion (V) is a widely used measure of dispersion and is defined as the ratio of the standard deviation to mean. In terms of mathematical notations, if x_1, x_2, \dots, x_n denote n real numbers with arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

and standard deviation

* Department of Mathematics, Himachal Pradesh University, Shimla (H.P.)

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2)$$

then

$$V = \frac{s}{\bar{x}}. \quad (3)$$

The coefficient V defined in (3) has the disadvantage of being affected very much by mean \bar{x} . We note that if we increase each value of a finite series by the value h , then the mean increases by h while the standard deviation remains the same. The formula (3) then gives different values of the coefficient of dispersion for the two series but by a simple observation we find that the two series must have the same amount of dispersion. In another case if the random variable takes both positive and negative values then \bar{x} may vanish and then (3) is not applicable. Further, there is no upper bound for V in general [1].

Moreover, we use a measure of dispersion to compare the variability of two series. A distribution having lesser coefficient of variation is said to be more consistent (or homogeneous) than the other. We consider an example to show that (3) does not always give the exactitude of variations in two distributions. Let X and Y be two discrete random variables taking values $\{1, 3, 11\}$ and $\{1, 9, 11\}$ respectively. Then we have

Table-1

Variable	Mean	Standard Deviation	Coefficient of Variation
X	5	4.32	0.86
Y	7	4.32	0.62

From Table-1, we see that the coefficient of variation of the distribution for Y is less than that for the distribution for X . Thus distribution for Y is more consistent than the distribution for X , which is not the case. Because by the symmetry of data, we see that two distributions are equally distributed and therefore should be equally consistent. The only difference is that the series X is dispersed towards the right while the series Y towards the left. These discrepancies in the Karl Pearson's coefficient of variation have led us to search for a measure for dispersion that does not suffer from the above stated deficiencies and must be equally applicable even when the random variable takes negative values.

Muirlwijk [2] shows that if $a \leq x_i \leq b, i = 1, 2, \dots, n$ then

$$s^2 \leq (b - a)(x - a). \quad (4)$$

The inequality (4) suggests us to propose the measure of dispersion in the following form

$$V = \frac{s}{\sqrt{(b-a)(x-a)}}. \quad (5)$$

where $a \leq x \leq b$. If $\bar{x} = a$ or $\bar{x} = b$, then $\bar{V} = 0$. We find that the coefficient of variation \bar{V} as defined by (5) has the following advantages over the coefficient of variation V as defined by (3).

(i) The coefficient of variation \bar{V} is bounded. By (4),

$$0 \leq \bar{V} \leq 1. \quad (6)$$

(ii) \bar{V} is independent of origin and scale. For, if x is a random variable in $[a, b]$ and

$$y = \frac{x - a}{\beta},$$

where α and β are real constants with $\beta \neq 0$, then

$$\bar{x} = \beta\bar{y} + \alpha \text{ and } s = \beta\sigma, \quad (7)$$

where \bar{y} is the mean and σ is the standard deviation of the random variable y . Then

it follows that *coefficient of variation of x and y are same.*

(iii) The coefficient of variation of V given by (3) 'suffers from disadvantage of being affected very much by the mean'. But such a disadvantage is not there in \bar{V} given by (5).

(iv) If we calculate \bar{V} for the case when mean and standard deviation are given in Table-1 then it comes out as same for the case of two random variables X and Y . Moreover it is found that if we have the random variable x taking values x_1, x_2, \dots, x_n with $x_1 \leq x_2 \leq \dots \leq x_n$ and variable y taking n values y_1, y_2, \dots, y_n such that $y_j = (x_1 + x_n) - x_{n-1+j}$ then a simple calculation shows that the values of \bar{V} for x and y are same while the values of V are different.

REFERENCES

M. Kendall and A. Stuart, The advanced theory of Statistics, Vol. 1, 4th Edition, Charles Griffin and Co. London, (1977).

J. Muilwijk, Note on a theorem of M.N. Murthy and V.K. Sethi, Sankhya Ser. B 28, 183, (1966).