Measures of Dispersion

Definition:

1. Two or more distributions may differ greatly in their dispersion, although their means may be the same, for e.g.:

$$67,67,67,67,67,67,67,67 \qquad \bar{x} = 67$$
$$43,43,50,55,66,90,91,97 \qquad \bar{x} = 67$$

2. By dispersion we mean the extent to which the values are spread out from the average. The measures used for computing the amount of dispersion in a distribution is known as 'measures of dispersion' or 'measures of variation'.
3. In the above distribution, the first distribution has zero dispersion, and the second distribution has a dispersion greater than the former. The dispersion cannot be less than zero.

Types of Measures of Dispersions:

Measures of dispersion are of two types:

(i) Measures of Absolute Dispersion, and
(ii) Measures of Relative Dispersion.

(i) **Measures of Absolute Dispersion:** The actual variation or dispersion determined by the Measures of Absolute Dispersion is called 'absolute dispersion'.

(ii) **Measures of Relative Dispersion:** The measures of absolute dispersion cannot be used to compare the variation of two or more series. For e.g., the SD of the height of students (in inches) cannot be compared with the SD of weights (in pounds). Even if the units are identical, for e.g., the comparison of height of men (in inches) and length of their noses (in inches). If the SD of heights of man is greater than the SD of their nose lengths, it does not mean that the degree of variability is greater in case of heights.

To compare the variation of two or more series, we need a measure of relative dispersion. It is defined as:

$$\text{Relative Dispersion} = \frac{\text{Absolute Dispersion}}{\text{Average}}$$

Types of Measures of Absolute Dispersion:

(a) The Range,
(b) The Quartile Deviation,
(c) The Mean Deviation, and
(d) The Standard Deviation.

(a) The Range:

1. The range is the simplest measure of dispersion. It is defined as the difference between the largest value and the smallest value in the data:

$$\text{Range} = X_{\max} - X_{\min}$$

2. For grouped data, the range is defined as the difference between the upper class boundary (UCB) of the highest class and the lower class boundary (LCB) of the lowest class.

(b) Quartile Deviation (QD):

1. It is also known as the Semi-Interquartile Range. The range is a poor measure of dispersion where extremely large values are present. The quartile deviation is defined half of the difference between the third and the first quartiles:

$$QD = Q_3 - Q_1$$

2. The difference between third and first quartiles is called the 'Inter-Quartile Range'.

(c) Mean Deviation (MD):

1. The MD is defined as the average of the deviations of the values from an average:

$$\text{MD from Mean} = \frac{\sum |x - \bar{x}|}{n}$$

It is also known as Mean Absolute Deviation.

2. MD from median is expressed as follows:

$$\text{MD from Median} = \frac{\sum |x - \tilde{x}|}{n}$$

3. For grouped data:

$$\text{MD}(\bar{x}) = \frac{\sum f|x - \bar{x}|}{\sum f}$$

$$\text{MD}(\tilde{x}) = \frac{\sum f|x - \tilde{x}|}{\sum f}$$

(d) Standard Deviation (SD):

1. The SD is defined as the positive Square root of the mean of the squared deviations of the values from their mean.
2. Thus, the SD of population of N values, $x_1, x_2, \ldots x_n$ is expressed as follows:

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$ ---------- Population Standard Deviation

3. In case of a frequency distribution with $x_1, x_2, \ldots, x_k$ as class marks, and $f_1, f_2, \ldots\ldots, f_k$ as the corresponding class frequencies, the SD is expressed as follows:

$$S = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$$

Alternate Method for Computing Standard Deviation:

1. If the values (or class marks) and the mean are not integral values, the computation of SD from its definition becomes labourious.
2. The shortcut alternate method for computing SD is:

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$ ---------- for ungrouped data (population SD)

$$S = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$$ ---------- for grouped data

3. If the values x are large, considerable time is served by taking deviations from x from an arbitrary value A. If D denotes deviations of x from A, i.e., D = x – A, then the SD can be expressed in another way:

$$\sigma = \sqrt{\frac{\sum D^2}{N} - \left(\frac{\sum D}{N}\right)^2}$$

4. Under coding method, the SD can be calculated as below:

$$S = h \times \sqrt{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2}$$

Where $u = \frac{x - A}{h} = \frac{D}{h}$

The Variance:

The variance is defined as the square of the SD, i.e., the mean of the squared deviations from mean:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$ ---------- for ungrouped data (population variance)

$$S^2 = \frac{\sum f(x - \bar{x})^2}{\sum f}$$ ---------- for grouped data

Sample Variance and Standard Deviation:

1. Variance of a sample of n values called sample variance, is expressed as below:

$$S^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

2. Standard deviation of sample of n values:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Alternate Method:

1. Variance:

$$S^2 = \frac{n \cdot \sum x^2 - \left(\sum x\right)^2}{n(n-1)}$$

2. Standard Deviation:

$$S = \sqrt{\frac{n \cdot \sum x^2 - \left(\sum x\right)^2}{n(n-1)}}$$

Properties of SD and Variance:

1. The SD or variance of a constant is zero. If x = a (a constant), SD(a) = 0 and var(a) = 0.
2. The SD and the variance are independent of origin, i.e., they remain unchanged when the values are increased or decreased by a constant:

$$SD(x + a) = SD(x); \; var(x + a) = var(x)$$
$$SD(x - a) = SD(x); \; var(x - a) = var\,(x)$$

3. When all the values are multiplied or divided by a constant the SD of these values is multiplied or divided by the constant and the variance is multiplied or divided by the square of the constant:

$$SD(ax) = a \times SD(x); \; var(ax) = a \times var(x)$$
$$SD(x/a) = (1/a) \times SD(x); \; var(x/a) = (1/a) \times var(x)$$

4. If two sets of data consisting of n_1 and n_2 have variances S_1^2 and S_2^2 respectively, the combined variance of both sets of data is expressed as follows:

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2} + \frac{n_1 + n_2}{\left(n_1 + n_2\right)} : \left(\bar{x}_1 - \bar{x}_2\right)$$

5. The variance of the sum or difference of two independent random variables is the sum of their respective variance. Thus, if x and y are independent random variables:

$$Var(x + y) = Var(x) + Var(y)$$
$$Var(x - y) = Var(x) - Var(y)$$

6. The variance has the minimal property. This means that the variance or the SD is minimum if and only if the deviation are taken from the mean. In other words:

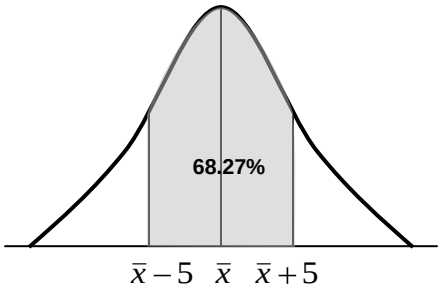
$$\frac{1}{n}\cdot\sum(x-a)^2 \text{ is a minimum when } a=\bar{x}$$

7. For normal distributions:

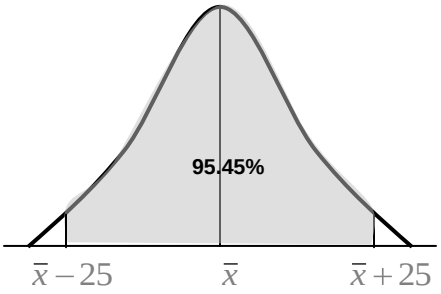
(i) the interval $\bar{x}-S$ to $\bar{x}+S$ includes 68.27% of the values,
(ii) the interval $\bar{x}-2S$ to $\bar{x}+2S$ includes 95.45% of the values, and
(iii) the interval $\bar{x}-3S$ to $\bar{x}+3S$ includes 99.73% of the values.

The above results also hold approximately for moderately skewed distributions.

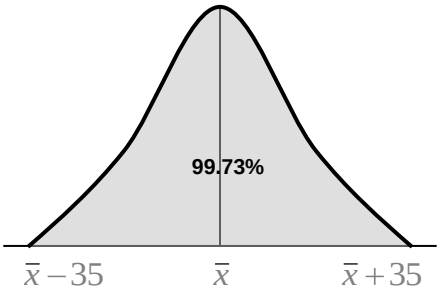
(i)



(ii)



(iii)



Characteristics of Measures of Dispersion:

(a) Range:

1. The range is simple to understand and easy to calculate because its value is determined by the two extreme items.
2. It is useful as a rough measure of variance.
3. Its value may be greatly changed if an extreme value (either lowest or highest) is withdrawn or a fresh value is added. It is a highly unstable measure of variation.
4. It gives no indication how the values within the two extremes are distributed.

(b) Quartile Deviation:

1. The QD is simple to understand and easy to calculate.
2. As a rough measure of variation, it is superior to the range because it is not affected by extreme values.
3. It is not capable of algebraic manipulation.
4. It is mainly used in situations where extreme values are thought to be un-representative.

(c) Mean Deviation:

1. The MD is simple to understand and to interpret.
2. It is affected by the value of every observation.
3. It is less affected by absolute deviations than the standard deviation.
4. It is not suited to further mathematical treatment. It is, therefore, not as logical as convenient measure of dispersion as the SD.

(d) Standard Deviation:

1. The SD is affected by the value of every observation.
2. The process of squaring the deviations before adding avoids the algebraic fallacy of disregarding signs.
3. In general, it is less affected by fluctuations of sampling than the other measures of dispersion.
4. It has a definite mathematical meaning and is perfectly adaptable to algebraic treatment.
5. It has great practical utility in sampling and statistical inference.
6. The SD is the best general purpose measure of dispersion and should be employed in all cases where a high degree of accuracy is required.

Example:

Class Boundaries	Frequency
9.5-19.5	5
19.5-29.5	8
29.5-39.5	13
39.5-49.5	19
49.5-59.5	23
59.5-69.5	15
69.5-79.5	7
79.5-89.5	5
89.5-99.5	3
99.5-109.5	2
Total	100

Calculate:

(a) Range
(b) Quartile deviation
(c) Mean deviation from mean
(d) Standard deviation
(e) Variance

Solution:

CB	f	CF	x	fx	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
9.5-19.5	5	5	14.5	72.5	-37.7	37.7	188.5	1421.29	7106.45
19.5-29.5	8	13	24.5	196	-27.7	27.7	221.6	767.29	6138.32
29.5-39.5	13	26	34.5	448.5	-17.7	17.7	230.1	313.29	4072.77
39.5-49.5	19	45	44.5	845.5	-7.7	7.7	146.3	59.29	1126.51
49.5-59.5	23	68	54.5	1253.5	2.3	2.3	52.9	5.29	121.67
59.5-69.5	15	83	64.5	967.5	12.3	12.3	184.5	151.29	2269.35
69.5-79.5	7	90	74.5	521.5	22.3	22.3	156.1	497.29	3481.03
79.5-89.5	5	95	84.5	422.5	32.3	32.3	161.5	1043.29	5216.45
89.5-99.5	3	98	94.5	283.5	42.3	42.3	126.9	1789.29	5367.87
99.5-109.5	2	100	104.5	209	52.3	52.3	104.6	2735.29	5470.58
Total	100			5220			1573		40371

(a) Range:

$$\text{Range} = X_{\max} - X_{\min} = 109.5 - 9.5 = 100$$

(b) Quartile Deviation:

$$Q_1 = \text{Value of } \frac{\sum f}{4} \text{ th item} = \frac{100}{4} = 25th\,item$$

$$Q_3 = \text{Value of } \frac{3\sum f}{4} \text{ th item} = \frac{3(100)}{4} = 75th\,item$$

$$Q_1 = l + \frac{h}{f}\left(\frac{\sum f}{4} - CF\right) = 29.5 + \frac{10}{13}\left(\frac{100}{4} - 13\right) = 38.73$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3\sum f}{4} - CF\right) = 59.5 + \frac{10}{15}\left(\frac{3(100)}{4} - 68\right) = 64.17$$

$$QD = Q_3 - Q_1 = 64.17 - 38.73 = 25.44$$

(c) Mean Deviation from Mean:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{5220}{100} = 52.2$$

$$\text{MD}(\bar{x}) = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{1573}{100} = 15.73$$

(d) Standard Deviation:

$$S = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}} = \sqrt{\frac{40371}{100}} = 20.09$$

(e) Variance:

$$S^2 = \frac{\sum f(x - \bar{x})^2}{\sum f} = \frac{40371}{100} = 403.71$$

Types of Measures of Relative Dispersions:

(a) Coefficient of Variation,
(b) Coefficient of Dispersion,
(c) Quartile Coefficient of Dispersion, and
(d) Mean Coefficient of Dispersion.

(a) Coefficient of Variation (CV):

1. Coefficient of variation was introduced by Karl Pearson. The CV expresses the SD as a percentage in terms of AM:

$$CV = \frac{S}{\bar{x}} \times 100 \text{ ---------- for sample data}$$

$$CV = \frac{\sigma}{\mu} \times 100 \text{ ---------- for population data}$$

2. It is frequently used in comparing dispersion of two or more series. It is also used as a criterion of consistent performance, the smaller the CV the more consistent is the performance.
3. The disadvantage of CV is that it fails to be useful when $\bar{x}$ is close to zero.
4. It is sometimes also referred to as 'coefficient of standard deviation'.
5. It is used to determine the stability or consistency of a data.

6. The higher the CV, the higher is instability or variability in data, and vice versa.

(b) Coefficient of Dispersion (CD):

If X_m and X_n are respectively the maximum and the minimum values in a set of data, then the coefficient of dispersion is defined as:

$$CD = \frac{X_m - X_n}{X_m + X_n} \times 100$$

(c) Coefficient of Quartile Deviation (CQD):

1. If Q_1 and Q_3 are given for a set of data, then $(Q_1 + Q_3)/2$ is a measure of central tendency or average of data. Then the measure of relative dispersion for quartile deviation is expressed as follows:

$$CQD = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

2. CQD may also be expressed in percentage.

(d) Mean Coefficient of Dispersion (CMD):

The relative measure for mean deviation is *'mean coefficient of dispersion'* or *'coefficient of mean deviation'*:

$$CMD = \frac{MD}{\bar{x}} \times 100 \text{ ---------- for arithmetic mean}$$

$$CMD = \frac{MD}{\tilde{x}} \times 100 \text{ ---------- for median}$$

Example:
(Take the previous example)

Calculate:

(a) Coefficient of Variation,
(b) Coefficient of Dispersion,
(c) Quartile Coefficient of Dispersion, and
(d) Mean Coefficient of Dispersion

Solution:

(a) Coefficient of Variation:

$$CV = \frac{S}{\bar{x}} \times 100 = \frac{20.09}{52.2} \times 100 = 38.49\%$$

(b) Coefficient of Dispersion:

$$CD = \frac{X_m - X_n}{X_m + X_n} \times 100 = \frac{109.5 - 9.5}{109.5 + 9.5} \times 100 = 84.03\%$$

(c) Quartile Coefficient of Dispersion:

$$CQD = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{(64.17 - 38.73)/2}{(64.17 + 38.73)/2} = \frac{12.72}{51.45} = 0.247 \text{ or } 24.7\%$$

(d) Mean Coefficient of Dispersion:

$$CMD = \frac{MD}{\bar{x}} \times 100 = \frac{15.73}{52.2} \times 100 = 30.13\%$$

Example:
During a soccer tournament, two players make the following series of goals:

Player 1	2	2	4	3	2	4	2	3
Player 2	1	2	5	5	5	2	1	1

Who is more consistent player?

Solution:

x	y	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$
2	1	-0.75	0.5625	-1.75	3.0625
2	2	-0.75	0.5625	-0.75	0.5625
4	5	1.25	1.5625	2.25	5.0625
3	5	0.25	0.0625	2.25	5.0625
2	5	-0.75	0.5625	2.25	5.0625
4	2	1.25	1.5625	-0.75	0.5625
2	1	-0.75	0.5625	-1.75	3.0625
3	1	0.25	0.0625	-1.75	3.0625
22	22		5.5		25.5

$$\bar{x} = \frac{\sum x}{n} = \frac{22}{8} = 2.75 \ ; \ \bar{y} = \frac{\sum y}{n} = \frac{22}{8} = 2.75$$

$$S_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{5.5}{8-1}} = 0.8864 \ ; \ S_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}} = \sqrt{\frac{25.5}{8-1}} = 1.9086$$

$$CV_x = \frac{S_x}{\bar{x}} \times 100 = \frac{0.8864}{2.75} \times 100 = 32.2\% \ ; \ CV_y = \frac{S_y}{\bar{y}} \times 100 = \frac{1.9086}{2.75} \times 100 = 69.4\%$$

Conclusion: The higher the CV, the higher the instability, and vice versa. From the above calculations, it is evident that Player 1 is more consistent than Player 2.

Standard Scores or Z-Scores:

Raw data can be converted into a special type of values by subtracting the mean from each value and then dividing by the SD of the data. These values are called 'standard scores' or 'z-scores' or 'values in SD units':

$$z = \frac{x - \bar{x}}{S} \text{ ---------- for sample data}$$

$$z = \frac{x - \mu}{\sigma} \text{ ---------- for population data}$$

Properties of Z-Score:

1. Z-scores are free of units.
2. The mean of z-scores is always zero.
3. The SD of z-scores is always one.
4. The distribution of z-scores looks exactly the same as the distribution of original data.

Example:

A student gets 82 marks in a final examination in Accounting; the mean is 75 marks with a standard deviation of 10 marks. In Economics, he gets 86 marks in the final examination on which the mean is 80 marks with a SD of 14 marks. Is his relative standing better in Accounting or Economics?

Solution:

Accounting	Economics
$\bar{x} = 75$ S = 10 x = 82 $z = \frac{x - \bar{x}}{S} = \frac{82 - 75}{10} = 0.7$	$\bar{x} = 80$ S = 14 x = 86 $z = \frac{x - \bar{x}}{S} = \frac{86 - 80}{14} = 0.43$

Conclusion: His marks in Accounting are 0.7 SD above the mean, while in Economics his marks are 0.43 SD above the mean. Therefore, his relative standing in Accounting is higher than Economics.

Chebyshev's Theorem:

1. A Russian mathematician P.L. Chebyshev has devised a rule called 'Chebyshev's Theorem' to determine the minimum proportion of values in intervals that are equidistant from mean.
2. The theorem states that for any data at least $\left(1-\frac{1}{k^2}\right)$ of the values must lie within k standard deviations on either side of the mean, where k is any constant number greater than 1.
3. In other words, the interval $\bar{x} \pm kS$ will contain at least $\left(1-\frac{1}{k^2}\right)$ of the values.

   For example:

   $\bar{x} \pm 2S$ will contain 75% of the values (k=2)
   $\bar{x} \pm 3S$ will contain 88.88% of the values (k=3)
   $\bar{x} \pm 2.4S$ will contain 82.64% of the values (k=2.4)

Limitations of Chebyshev's Theorem:

1. Proportions of values are given only for intervals which are equidistant from mean, that is the mean should always be the mid-point of the interval.
2. Minimum proportion is specified rather than exact or approximate value of the proportion.
3. Proportions for values of k less than or equal to one cannot be determined.

Example:

Two populations have the same mean $\mu = 140$. Their SDs are $\sigma_1 = 10$ or $\sigma_2 = 3$. Find the percentages of the values that must lie between 125 and 155.

Solution:

Population 1		**Population 2**	
$\mu_1 = 140$ $\sigma_1 = 10$ $\mu \pm k\sigma$		$\mu_2 = 140$ $\sigma_2 = 3$ $\mu \pm k\sigma$	
$140 - k \cdot 10 = 125$ $k \cdot 10 = 15$ $k = \frac{15}{10} = 1.5$	$140 + k \cdot 10 = 155$ $k \cdot 10 = 15$ $k = \frac{15}{10} = 1.5$	$140 - k \cdot 3 = 125$ $k \cdot 3 = 15$ $k = \frac{15}{3} = 5$	$140 - k \cdot 3 = 155$ $k \cdot 3 = 15$ $k = \frac{15}{3} = 5$

Therefore 125 to 155 will contain at least:	Therefore 125 to 155 will contain at least:
$= \left(1 - 1/k^2\right) \times 100$	$= \left(1 - 1/k^2\right) \times 100$
$= \left(1 - 1/(1.5)^2\right) \times 100$	$= \left(1 - 1/(5)^2\right) \times 100$
$= 55.6\%$ of the values	$= 96\%$ of the values

Normal Distribution:

1. Three mathematicians, namely, P. Laplace, A. De Moivre and K.F. Gauss have independently developed a law which gives the proportion of values that lie in specific intervals of a special type of symmetrical distribution called 'Normal Distribution'.
2. The mathematical form of a normal distribution is complicated and difficult to use frequently. Tables have constructed to make the application of normal law simple, known as 'tables of areas under normal curve' or 'normal area tables'.
3. Whenever the frequency curve is bell shaped or symmetrical, the distribution (or curve) can be assumed approximately normal and hence normal law can be applied.

Interval	Percentage of Values
$\mu \pm \sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

Linear Transformation of a Variable:

1. Let $\bar{x}$ and S_x be the mean and SD of a variable x.
2. Let the variable x multiplied by a constant number and a constant number added to the product giving a new variable y.
3. Then the variable x is said to be linearly transformed to the variable y and the process is called a 'linear transformation of x to y'.
4. Symbolically $y = k + h \cdot x$ is a linear transformation where k and h are any constant numbers.
5. The mean and SD of the transformed variable y may be expressed in terms of the mean and SD of the variable x by the following relations:

$$\bar{y} = k + h \cdot \bar{x}$$

$$S_y = h \cdot S_x$$

6. It should be noted here that the z-score is a linear transformation of a variable x such that:

$$k = -\frac{\bar{x}}{S} \text{ and } h = \frac{1}{S}$$

$$\text{Since } z = \frac{x - \bar{x}}{S} \text{ or } z = \frac{x}{S} - \frac{\bar{x}}{S}$$

$$z = -\frac{\bar{x}}{S} + \frac{1}{S} \cdot x$$

$$z = k + h \cdot x$$

Example:

Given: $\bar{x} = 25$ and $S_x = 5$.

Determine the mean and standard deviation of the following transformations of x:

(i) $y = 10 + x$
(ii) $3x = 2y$

Solution:

(i) $y = 10 + x$:

Rules:
SD(x + a) = SD(x)
SD(ax) = a × SD(x)

$\bar{y} = 10 + \bar{x}$
$\bar{y} = 10 + 25 = 35$
$S_y = S_x = 4$

(ii) $3x = 2y$:

Rules:
SD(x + a) = SD(x)
SD(ax) = a × SD(x)

$3\bar{x} = 2\bar{y}$

$3(25) = 2\bar{y}$

$75 = 2\bar{y}$

$\bar{y} = \frac{75}{2} = 37.5$

$3S_x = 2S_y$

$3(5) = 2S_y$

$2S_y = 15$

$S_y = 7.5$