MEASURES OF DISPERSION

Introduction

Dispersion is the second important characteristic of a frequency distribution. Two distributions may have the same mean, median and mode, but the variation of the individual observations of the two distributions may differ. For example, the daily wages in Taka of seven workers of two factories are as follows:

Wages of factory : I	142	143	150	150	153	155	157
Wages of factory : II	122	140	150	150	154	159	175

Here it is seen that the mean, median and modal wages of both the factories are same. It is Tk. 150. That is both the distributions have the same measures of location. But the variation of the observations are not same. The range of the wage structure of the first factory is Tk. $15 = 157 - 142$ only whereas the range of the second factory is Tk. $53 = 175 - 122$. That is the wage structure of the factory I is more compact than the wage structure of the factory II. The variability or dispersion of the individual values of factory 1 is less than the factory 2. Suppose the depths of a river at 5 different points are 4, 3, 2, 3, 8 feet. The average depth of the river at different points is 4ft; it would not give you the guarantee to cross the river safely. To cross the river safely, you have to know the maximum depth of the river. This means measures of location have failed to measure the variability of a set of data. The need for a measure of dispersion in addition to a measure of location is necessary. Small dispersion indicates high uniformity of the observations in the distribution.

Dispersion tells us how compactly the individual values are distributed around the central values. Dispersion of a single variable might not bear that much meaning, while comparison of dispersion of two sets of variables is more useful for taking decision.

According to Books and Dicks "Dispersion or spread is the degree of the scatter or variation of the variables about a central value".

Definition. Dispersion measures the variability of a set of observations among themselves or about some central values.

Purposes of Dispersion

Measure of dispersion is needed for four basic purposes:

i) To determine the reliability of an average.
ii) To serve as a basis for the control of the variability.
iii) To compare two or more series with regard to their variability.
iv) To facilitate the computation of other statistical measures.

Properties of a Good Measures of Dispersion

The properties of a good measure of dispersion are the same as those of a good measure of central tendency. They are as follows:

i) It should be simple to understand,
ii) It should be easy to compute,
iii) It should be rigidly defined,
iv) It should be based on all the observations,
v) It should have sampling stability,
vi) It should be suitable for further algebraic treatment, and
vii) It should not be affected by extreme observations.

Measures of Dispersion

The numerical values by which we measure the dispersion or variability of a data set or a frequency distribution are called measures of dispersion.

There are two kinds of measures of dispersion. They are

i) Absolute measures of dispersion and ii) Relative measures of dispersion.

Important absolute measures of dispersion are

i) Range, ii) Quartile deviation, iii) Mean deviation, and iv) Variance and Standard deviation.

The relative measures of dispersion are i) Coefficient of range, ii) Coefficient of quartile deviation, iii) Coefficient of mean deviation, and iv) Coefficient of variation.

It is seen that for each absolute measure of dispersion, there is a relative measure of dispersion. Absolute measures are expressed in the same unit in which the original variables are measured. For example, any absolute measure of weight may be measured as kilogram or pound, height as meter or inch, price as taka or dollar etc. On the other hand, relative measures are pure number and independent of the unit of measurement and express in percentage.

Hence relative measures are better than absolute measures to compare the variability of two sets of observations or distributions measured in different units. Now we shall discuss them.

Range

Range is the difference between the largest and smallest observations in a set of data. Symbolically, Range $= X_L - X_S$. Here, X_L = largest observation and X_S = smallest observation.

Coefficient of range. The relative measure corresponding to range, called the coefficient of range is computed by the following formula:

$$\text{Coefficient of range} = \frac{X_L - X_S}{X_L + X_S} \times 100 .$$

Merits and limitations. The following are the merits and limitations of range:

Merits

i) The range measures the total spread in a data set.
ii) It is rigidly defined.
iii) It is the simplest measure of dispersion and easiest to compute.
iv) It takes minimum time to compute.
v) It is based on only maximum and minimum values of a data set.

Limitations

i) It is not based on all the values of a set of data.
ii) It is affected by sampling fluctuation.
iii) It cannot be computed in case of open-end distribution.
iv) It is highly affected by extreme values.

Uses of Range. Although range measures the total spread of a set of observations, its uses are prominent in following fields:

i) **Quality Control.** It is widely used in production process to control the quality of the products.
ii) **Share Market.** It is widely used to study the variations in the prices of stocks and shares and other commodities.
iii) **Weather forecasts.** The meteorological department uses the range to determine the difference between the minimum and maximum temperature, which is a very useful index for people to know the limits of temperature in a particular season. Also maximum and minimum values of

other climatic factors such as rainfall, humidity, wind velocity etc. are very important from the metrological point of view.

Inter - Quartile Range and Quartile Deviation

Another measure similar to range is the inter-quartile range.

Definition. The range which includes the middle 50% of the observations is called inter-quartile range. In other words, it is the difference between the third quartile and the first quartile. Symbolically, if Q_3 and Q_1 are third and first quartiles of a data set respectively then inter-quartile range is:

$$\text{Inter - Quartile Range (IQR)} = Q_3 - Q_1 .$$

Quartile deviation. Half of the inter-quartile range is called quartile deviation. It is also sometimes known as semi-inter-quartile range (SIQR). Symbolically, if Q_3 and Q_1 are third and first quartiles of a data set, quartile deviation, denoted by Q.D. is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} .$$

Quartile deviation is better than range as it measures the variation of the middle 50% of the observations. Small quartile deviation means high uniformity or small variation of the central 50% observations, whereas a large quartile deviation means large variation among the central observations.

Coefficient of quartile deviation. Quartile deviation is an absolute measure of variation. The relative measure corresponding to this measure, called the coefficient of quartile deviation, is defined by

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 .$$

It is used to compare the variation of two distributions measured in different units of measurements.

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate quartile deviation and coefficient of quartile deviation.

Solution. Quartile deviation and coefficient of quartile deviations are computed by the formulas:

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} \quad \text{and} \quad \text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100.$$

First quartile, Q_1, is the $n/4$ ordered observation = 12.5th ordered observation. It lies in the class interval 80 – 105.

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1$$

Here $L = 80$, $n/4 = 12.5$, $F_1 = 7$, $f_1 = 6$ and $c_1 = 25$. Hence,

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1 = 80 + \frac{12.5 - 7}{6} \times 25$$

$= 80 + \frac{5.5 \times 25}{6} = 80 + 22.92 = 102.92$ hours per month.

Third quartile, Q_3 is the 3n/4th ordered observation = 37.5th ordered observation. It lies in the class 155 - 180. Third quartile is:

$$Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3$$

Here, $L = 155$, $3n/4 = 37.5$, $F_3 = 34$, $f_3 = 11$, $c_3 = 25$.

$$Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3 = 155 + \frac{37.5 - 34}{11} \times 25 = 155 + \frac{3.5 \times 25}{11}$$

$= 155 + 7.95 = 162.95$ hours per month.

Hence Q.D. $= \frac{Q_3 - Q_1}{2} = \frac{162.95 - 102.92}{2} = \frac{60.03}{2} = 30.015$ hours per month.

Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 = \frac{162.95 - 102.92}{162.95 + 102.92} \times 100 = \frac{60.03}{265.87} \times 100 = 22.58\%$.

Merits and demerits of quartile deviation. The quartile deviation of a set of observations possesses following merits:

i) It is superior to range as a measure of variation.
ii) It is useful in case of open-end distribution.
iii) It is not affected by the presence of extreme values.
iv) It is useful in case of highly skewed distribution.

While the Q.D. is liable for following limitations:

i) It is not based on all the observations.
ii) Quartile deviation ignores the first 25% and the last 25% observations.
iii) It is not capable of mathematical manipulation.
iv) Its value is very much affected by sampling fluctuations.
v) It is not a good measure of dispersion since it depends on two positional measures.

Mean Deviation

Range and quartile deviation do not measure the scatterness of the observations about any central values. However, to study the formation of a distribution we should take the deviation of the observations from an average. Mean deviation is based on such kind of deviation.

Mean Deviation is obtained by calculating the absolute deviations of each observation from mean or median or mode and then averaging these deviations by taking their arithmetic mean. So we can define mean deviation in three ways, namely,

i) Mean deviation about mean; ii)Mean deviation about median; and
iii) Mean deviation about mode.

Mean deviation for ungrouped data.

Mean Deviation. Suppose $x_1, x_2, \ldots\ldots\ldots, x_n$ are n observations of a data set, and $\bar{x}$ is the mean, then mean deviation (M.D.) about mean is defined by

$$\text{M.D.}(\bar{x}) = \frac{\Sigma|x - \bar{x}|}{n}$$

Similarly, mean deviation about median and mean deviation about mode are defined by

Comment. Since the grade-point averages of all the students are more than 4 and the standard deviation is only 0.59, the quality of the students are very good (grade points of the student are close to each other and around 4).

Sample variance by the second computational formula:

$$s^2 = \frac{1}{n-1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] = \frac{1}{5}\left[115.27 - \frac{(26.1)^2}{6}\right]$$

$$= \frac{1}{5}\left[115.27 - \frac{681.21}{6}\right] = \frac{1}{5}[115.27 - 113.535] = \frac{1.735}{5} = 0.347.$$

Notation	
Sample	Population
n : number of observations in the sample	N : number of observations in the population
s^2 : sample variance	σ^2 : population variance
$s = \sqrt{s^2}$: sample standard deviation	$\sigma = \sqrt{\sigma^2}$: population standard deviation

Sample variance for grouped data. Let $x_1, x_2, \ldots\ldots, x_k$ be k values of a variable or k mid points of k classes with corresponding frequencies $f_1, f_2, \ldots\ldots, f_k$ then the sample variance is defined by

$$s^2 = \frac{\Sigma f(x-\bar{x})^2}{n-1}; \qquad \text{where } n = \Sigma f.$$

For convenience and simplicity most of the textbooks use the divisor n in place of n-1 and they use the following formula for sample variance:

$$s^2 = \frac{\Sigma f(x-\bar{x})^2}{n};$$

However, the computing formula for sample variance is

$$s^2 = \frac{1}{n-1}\left[\Sigma fx^2 - \frac{(\Sigma fx)^2}{n}\right] \quad \text{or,} \quad s^2 = \frac{n\Sigma fx^2 - (\Sigma fx)^2}{n(n-1)}.$$

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate variance and standard deviation of the frequency distribution.

Solution. Calculation of variance and standard deviation

Class interval	Mid-point : x	Frequency : f	fx	fx^2
30-55	42.5	3	127.5	5418.75
55-80	67.5	4	270.0	18225.00
80-105	92.5	6	555.0	51337.50
105-130	117.5	9	1057.5	124256.25
130-155	142.5	12	1710.0	243675.00
155-180	167.5	11	1842.5	308618.75
180-205	192.5	5	9672.5	185281.25
Total		50	$\Sigma fx = 6525$	$\Sigma fx^2 = 936812.5$

$$s^2 = \frac{1}{n-1}\left[\Sigma fx^2 - \frac{(\Sigma fx)^2}{n}\right] = \frac{1}{49}\left[936812.50 - \frac{(6525)^2}{50}\right]$$

$$= \frac{1}{49}[936812.50 - 851512.50] = \frac{85300}{49} = 1740.82 = 1740.82$$

Hence, $s = \sqrt{1740.82} = 41.72$ hours.

Calculation of variance and standard deviation by short cut method. Suppose $x_1, x_2, \ldots\ldots, x_k$ are k values of a variable or k mid points of k classes with corresponding frequencies $f_1, f_2, \ldots\ldots, f_k$ then the sample variance is:

$$s^2 = \frac{\Sigma f(x-\bar{x})^2}{n-1}; \qquad \text{where } n = \Sigma f$$

Let, $d = \frac{x-A}{i}$; then $x = A + id$ and $\bar{x} = A + i\bar{d}$.

Here A is called assumed mean which is usually taken as middle value of x or the value of x which has the highest frequency to get the maximum benefit of calculation and i is the width of the class interval.

Substituting the value of $(x-\bar{x})$ in the formula of s^2, we have:

$$s^2 = \frac{\Sigma f(x-\bar{x})^2}{n-1} = \frac{\Sigma f[i(d-\bar{d})]^2}{n-1} = \frac{\Sigma f(d-\bar{d})^2}{n-1} \times i^2 = \frac{1}{n-1}\left[\Sigma fd^2 - \frac{(\Sigma fd)^2}{n}\right] \times i^2$$

Then standard deviation is computed by the following formula:

$$s = \sqrt{\frac{1}{n-1}\left[\Sigma fd^2 - \frac{(\Sigma fd)^2}{n}\right]} \times i.$$

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory.

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate variance and standard deviation of the frequency distribution by the short cut method.

Solution. Let $A = 117.5$ and $i = 25$; then $d = \frac{x-A}{i} = \frac{x-117.5}{25}$.

Calculation of variance and standard deviation

Class interval	Mid-point : x	Frequency : f	$d = \frac{x-117.5}{25}$	fd	fd^2
30-55	42.5	3	-3	-9	27
55-80	67.5	4	-2	-8	16
80-105	92.5	6	-1	-6	6
105-130	117.5	9	0	0	0
130-155	142.5	12	1	12	12
155-180	167.5	11	2	22	44
180-205	192.5	5	3	15	45
Total		50		$\Sigma fd = 26$	$\Sigma fd^2 = 150$

$$\text{Variance} = s^2 = \frac{1}{n-1}\left[\Sigma fd^2 - \frac{(\Sigma fd)^2}{n}\right] \times i^2 = \frac{1}{49}\left[150 - \frac{(26)^2}{50}\right] \times (25)^2$$

$$= \frac{1}{49}[150 - 13.52] \times 625 = \frac{136.48 \times 625}{49} = 1740.82.$$

Hence the standard deviation, $s = \sqrt{1740.82} = 41.72$.

From the computation point of view, it is the easiest method of calculating variance as well as standard deviation.

Merits and demerits of standard deviation. Merits and demerits of standard deviation are as follows:

Merits

i) It is rigidly defined.
ii) It is based on all observations of the distribution.
iii) It is amenable to algebraic treatment.
iv) It is less affected by sampling fluctuation.
v) It is possible to calculate the combined standard deviation of two or more groups.

Demerits and Limitations

i) As compared to other measures it is difficult to compute.
ii) It is affected by the extreme values.
iii) It is not useful to compare two sets of data when the observations are measured in different units.

Some comments on standard deviation. It is the best absolute measure of dispersion. Some important characteristics of standard deviation are:

1. The value s is always greater than or equal to zero.
2. The larger the value of s, the greater the variability of the data set.
3. If s is equal to zero, all the observations must have the same value. That is there is no variability among the observations in the data set.
4. The original variable and the standard deviation have the same unit of measurement.

General comments on the absolute measures of dispersion:

1. The more spread out or dispersed the data are, the larger will be the range, the quartile deviation, the variance and the standard deviation.
2. The more concentrated or homogenous the data are, the smaller will be the range, the quartile deviation, the variance and the standard deviation.
3. If the observations are all the same (so that there is no variation in the data), the range, the quartile deviation, the variance and the standard deviation will be all zero.
4. None of the measures of dispersion can be negative.

Empirical Relations among the measures of Dispersion. For a symmetrical and moderately skewed distribution there exist relationships among the three commonly used measures of dispersion. The quartile deviation (Q.D.) is the smallest, following the mean deviation (M.D.) and the standard deviation (S.D.) is the largest. Following are the relationships among themselves:

$$\text{Q.D.} = \frac{2}{3}\sigma \; ; \qquad \text{M.D.} = \frac{4}{5}\sigma \; ; \qquad \text{and} \qquad \text{Q.D.} = \frac{5}{6}\text{M.D.}$$

They are useful in estimating one measure of dispersion when another is known or in verifying the consistency of the calculated values roughly. If the computed σ differs very widely from its value estimated from Q.D. or M.D. either an error has been made or the distribution differs considerably from symmetry.

Another comparison may be made of the proportion of observations that are typically included within the range of one Q.D., M.D. or S.D. measured both above and below the mean. For a normal distribution:

$\bar{x} \pm$ Q.D. , includes 50% of the observations.

$\bar{x} \pm$ M.D. , includes 57.51% of the observations.

$\bar{x} \pm$ S.D. , includes 68.27% of the observations.

Combined variance and standard deviation of two populations: Suppose we have two populations having sizes N_1 and N_2 observations with variances σ_1^2 and σ_2^2 respectively, and μ_1 and μ_2 are their respective means and μ_{12} is the combined mean of the two populations taken together, then the combined variance σ_{12}^2 of the two populations is

$$\sigma_{12}^2 = \frac{(N_1\sigma_1^2 + N_2\sigma_2^2) + (N_1 d_1^2 + N_2 d_2^2)}{N_1 + N_2}; \text{ where } d_1 = \mu_{12} - \mu_1, d_2 = \mu_{12} - \mu_2.$$

The combined standard deviation of two populations is $\sigma_{12} = \sqrt{\sigma_{12}^2}$.

Combined variance and standard deviation of three populations: Similarly, the combined variance of three populations is

$$\sigma_{123}^2 = \frac{(N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2) + (N_1 d_1^2 + N_2 d_2^2 + N_3 d_3^2)}{N_1 + N_2 + N_3}$$

Here σ_1^2, σ_2^2 and σ_3^2 are the variances of the sets 1, 2 and 3 and N_1, N_2 and N_3 are the items containing the sets 1, 2 and 3 respectively. And $d_1 = \mu_{123} - \mu_1$, $d_2 = \mu_{123} - \mu_2$ and $d_3 = \mu_{123} - \mu_3$. The combined standard deviation of three populations is

$$\sigma_{123} = \sqrt{\sigma_{123}^2}.$$

Combined variance and standard deviation of two samples: Suppose we have two samples containing n_1 and n_2 observations with variances s_1^2 and s_2^2 respectively. Suppose $\bar{x}_1$ and $\bar{x}_2$ are their respective means and $\bar{x}_{12}$ is the combined mean of the two sets taken together, then the formula for finding combined variance s_{12}^2 is

$$s_{12}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 -)s_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2 - 2}; \text{ where, } d_1 = \bar{x}_{12} - \bar{x}_1 \text{ and } d_2 = \bar{x}_{12} - \bar{x}_2.$$

The combined standard deviation is the positive square root of the combined variance. That is, $s_{12} = \sqrt{s_{12}^2}$.

The formula for combined variance in case of three sets of observations is

$$s_{123}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + n_1 d_1^2 + n_2 d_2^2 + n_3 d_3^2}{n_1 + n_2 - 3}$$

Here s_1^2, s_2^2 and s_3^2 are the variances of the sets 1, 2 and 3 and n_1, n_2 and n_3 are the items containing the sets 1, 2 and 3 respectively. And $d_1 = \bar{x}_{123} - \bar{x}_1$, $d_2 = \bar{x}_{123} - \bar{x}_2$ and $d_3 = \bar{x}_{123} - \bar{x}_3$.

The combined standard deviation is the positive square root of the combined variance. That is, $s_{123} = \sqrt{s_{1,3}^{\ 2}}$.

Coefficient of Variation

Standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the coefficient of variation. This measure developed by Karl Pearson is the most commonly used measure of relative dispersion.

Definition. If μ is the mean and σ is the standard deviation of a population data set, then coefficient of variation denoted by C.V. is defined by

$$C.V. = \frac{\sigma}{\mu} \times 100.$$

If $\bar{x}$ is the mean and s is the standard deviation of a sample data set, then coefficient of variation is defined by

$$C.V. = \frac{s}{\bar{x}} \times 100.$$

It is a pure number and expressed as a percentage. It is useful in comparing the variability of two or more sets of data, especially if they are expressed in different units of measurement. Since it is a ratio, the units of measurement have no significance. Moreover, coefficient of variation depends on the best measure of central tendency and the best measure of dispersion. In these reasons, coefficient of variation is better than standard deviation as a measure of dispersion. Overall, coefficient of variation is the best measure of dispersion among all the absolute and relative measures.

Comment. Coefficient of variation is the best measure of dispersion.

Example 6.9.1. The grade point averages obtained by randomly selected 6 students in their H.S.C. examination are as follows: 4.9, 4.1, 4.4, 3.3, 4.6 and 4.8. Find mean, standard deviation and coefficient of variation.

Solution. We know from the problem 6.8.3 that the standard deviation of the data set is 0.59 and $\Sigma x = 26.1$. Then, $\bar{x} = \frac{26.1}{6} = 4.35$.

Hence, $C.V. = \frac{s}{\bar{x}} \times 100 = \frac{0.59}{4.35} \times 100 = 13.56.$

Among two or more data sets, the set whose coefficient of variation is greater is said to be more variability or conversely less consistent, less uniform, less stable or less homogenous. On the other hand, the series for which coefficient of variation is less is said to be less variability or more consistent, more uniform, more stable or more homogenous.

Example Lives of two models of refrigerators in a recent survey were found as follows:

Life (no. of years)	Model A	Model B
0-2	5	2
2-4	16	7
4-6	13	12
6-8	7	19
8-10	5	9
10-12	4	1

i) What is the average life of each model of these refrigerators?

ii) Which of the two models shows more uniformity?

iii) A person wants to buy a new refrigerator, which one will he prefer? N.U.2005

Solution. For finding the average lifetimes, we have to compute arithmetic mean and for determining the model which has greater uniformity, then compute and compare the coefficient of variations.

Class interval	Mid-points : x	Model A			Model B		
		f	fx	fx^2	f	fx	fx^2
0-2	1	5	5	5	2	2	2
2-4	3	16	48	144	7	21	63
4-6	5	13	65	325	12	60	300
6-8	7	7	49	343	19	133	931
8-10	9	5	35	405	9	81	729
10-12	11	4	44	484	1	11	121
Total		50	256	1706	50	308	2146

Computations of mean, variances and co-efficient of variations of lifetimes for two models are shown below:

	Model A	Model B
Arithmetic Mean,	$\bar{x}_A = \frac{256}{50} = 5.12$ years	$\bar{x}_B = \frac{304}{50} = 6.16$ years
	$s_A^2 = \frac{1}{49}\left[1706 - \frac{(256)^2}{50}\right]$	$s_B^2 = \frac{1}{49}\left[2146 - \frac{(308)^2}{50}\right]$
	$= \frac{1}{49}[1706 - 1310.72]$	$= \frac{1}{49}[2146 - 1897.28]$
	$= \frac{395.28}{49} = 8.07$.	$= \frac{248.72}{49} = 5.08$.
	$s_A = \sqrt{8.07} = 2.84$ years.	$s_B = \sqrt{5.08} = 2.25$ years.
	C.V. (A) $= \frac{2.84}{5.12} \times 100 = 55.47\%$.	C.V.(B) $= \frac{2.25}{6.16} \times 100 = 36.53\%$.

i) Average life times of refrigerators of Model A is 5.12 years, while of Model B is 6.16 years.
ii) Since coefficient of variation is less for Model B, hence refrigerators of Model B show greater uniformity as per the lifetime of the refrigerators.
iii) Due to the greater uniformity in lifetime, the person will prefer Model B.

Example The following are some of the particulars of the distribution of weights of boys and girls in a class.

	Boys	Girls
Number :	65	35
Mean weight :	60kgs	45kgs
Standard deviation :	4kgs	2kgs

The weights of which distribution is more homogenous?

Solution. We can consider the data from two populations. To compare the variability of the weights of two distributions, we will have to find the coefficient of variations of the two distributions.

Boys	Girls
C.V. $= \frac{\sigma}{\mu} \times 100 = \frac{4}{60} \times 100 = 6.67\%$.	C.V. $= \frac{\sigma}{\mu} \times 100 = \frac{2}{45} \times 100 = 4.44\%$.

Comment. Since the coefficient of variation is less for girls, hence the weights of the girls are less variability and more homogenous.

Advantage of coefficient of variation over standard deviation. Standard deviation and the original variable have the same unit of measurements. So standard deviation cannot be used to compare the variability of two or more distributions measured in different units. But coefficient of variation is a pure number. That is coefficient of variation is independent of the unit of measurement of the original variables. Therefore, coefficient of variation can be successfully used to compare the variability of two or more distributions in different units of measurement. Hence coefficient of variation is sometimes better than standard deviation as a measure of dispersion.

Some Elementary Theorem and Examples

Theorem Variance as well as standard deviation is independent of the shift of origin but depends on change of scale.

Proof. Suppose $x_1, x_2, \ldots\ldots\ldots, x_n$ are n values of a sample with mean $\bar{x}$, then variance is define by

$$s_x^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

Then the sample standard deviation is

$$s_x = \sqrt{\frac{\Sigma(x-\bar{x})}{n-1}}.$$

Let, $u = \frac{x-A}{h}$. Then, $x = A + hu$ and $\bar{x} = A + h\bar{u}$.

Now, by putting the values of x and $\bar{x}$ in (6.11.1), we have

$$s_x^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{\Sigma(A+hu-A-h\bar{u})^2}{n-1} = h^2\frac{\Sigma(u-\bar{u})^2}{n-1} = h^2 s_u^2 .$$

That is variance of x is h^2 times variance of u. It is seen that variance of x depends on h but not on A. Hence variance is independent of the shift of origin but depends on scale. Moreover, standard deviation of x is $s_x = hs_u$.

This shows that standard deviation also independent of the shift of origin but depends on scale.

Theorem For two numbers standard deviation is the half of the range.

Proof. Let x_1 and x_2 are two quantities such that $x_1 > x_2$.

Then range is: $R = x_1 - x_2$; since $x_1 > x_2$.

Here, $\mu = \frac{x_1+x_2}{2}$. Then

$$\text{Variance} = \sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2}{2} = \frac{\left(x_1-\frac{x_1+x_2}{2}\right)^2+\left(x_2-\frac{x_1+x_2}{2}\right)^2}{2} = \frac{(x_1-x_2)^2}{4}$$

$$\text{Standard deviation} = \sigma = \frac{x_1-x_2}{2} = \frac{R}{2}.$$

Hence for two quantities standard deviation is the half of the range.

From here it follows that for two positive quantities mean is always greater than standard deviation, since $\frac{x_1+x_2}{2} \geq \frac{x_1-x_2}{2}$.

Equality holds when any one of them is equal to zero.

Theorem The variance of the first n natural numbers is: $\frac{n^2-1}{12}$.

Proof. Let x be a variable whose values are 1, 2, 3, ..., n.

Then, Mean $= \mu = \frac{1+2+3+.....+n}{n} = \frac{n(n+1)/2}{n} = \frac{n+1}{2}$.

Variance $= \sigma^2 = \frac{1}{n}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]$. Here, $\Sigma x^2 = 1^2+2^2+3^2.....+n^2 = \frac{n(n+1)(2n+1}{6}$. Then

variance $= \sigma^2 = \frac{1}{n}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] = \frac{1}{n}\left[\frac{n(n+1)(2n+1}{6} - \frac{n(n+1)^2}{4}\right]$

$$= \frac{n(n+1)}{n}\left[\frac{2n+1}{6} - \frac{n+1}{4}\right] = (n+1)\left[\frac{4n+2-3n-3}{12}\right] = \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12}.$$

Hence the standard deviation of the first natural numbers is

$$\sigma = \sqrt{\frac{n^2-1}{12}}.$$