Measures of Dispersion or Spread or Variation

Measures of Dispersion: Sometimes when two or more different data sets are to be compared using measure of central tendency or averages, we get the same result.

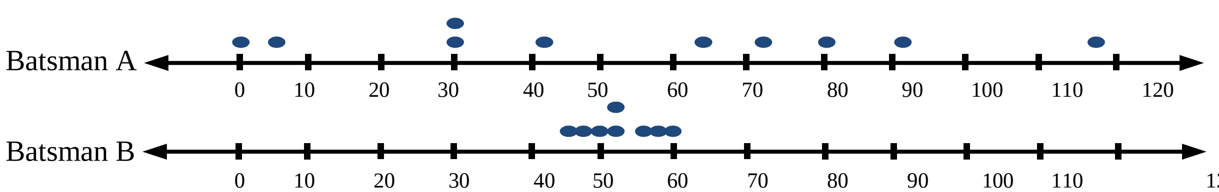
Consider the runs scored by two batsmen in their last ten matches as follows:

Batsman A: 30, 91, 0, 64, 42, 80, 30, 5, 117, 71
Batsman B: 53, 46, 48, 50, 53, 53, 58, 60, 57, 52

Clearly, mean of the runs scored by both the batsmen A and B is same i.e. 53

Can we say that the performance of two players is same? Clearly No, because the variability in the scores of batsman A is from 0 to 117, whereas, the variability of the runs scored by batsman B is from 46 to 60.

Let us now plot the above scores as dots on a number line. We find the following diagrams:



We can see that the dots corresponding to batsman B are close to each other and is clustering around the measure of central tendency (mean), while those corresponding to batsman A are scattered or more spread out. Thus, the measures of central tendency are not sufficient to give complete information about a given data. In such a situation the comparison becomes very difficult. We therefore, need some additional information for comparison, concerning with, how the data is dispersed about (more spread out) the average. This can be done by measuring the dispersion. Like „measures of central tendency‟ we want to have a single number to describe variability. This single number is called a **‘measure of dispersion’**.

Dispersion: “The variability (spread) that exists between the value of a data is called dispersion”
OR,
“The extent to which the observations are spread around an average is called dispersion or scatter”.

> *"Dispersion is the measure of the variation of the items."*
> —A.L. Bowley
>
> *"Dispersion is a measure of the extent to which the individual items vary."* —L.R. Connor
>
> *"Dispersion or spread is the degree of the scatter or variation of the variables about a central value."*
> —B.C. Brooks and W.F.L. Dicks
>
> *"The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data."* —Spiegal

Purpose of measure of dispersion

- Measure of dispersion is important for the following purpose.
- To determine the reliability of an average.
- To compare the variability.
- To compare two or more series with regard to their variability.
- To facilitate the use of other statistical measures.
- It is one of the most important quantities used to characterize a frequency distribution.

Characteristics of a good measure of dispersion- The following are the characteristics of an ideal measure of variation or dispersion:

I. It should be easy to understand.
II. It should be easy to calculate.
III. It should be based upon all observations.
IV. It should be rigidly defined.
V. It should not be unduly affected by extreme values.
VI. It should be suitable for further algebraic treatment.
VII. It should be less affected by sampling fluctuation.

As we know that, there are quite a few ways of measuring the central tendency of a data set i.e. A.M, G.M, H.M, Mode and Median. Similarly, we have different ways of measuring and comparing the dispersion of the distribution(s). There are two important types of measures of dispersion.

 Types of Measures of Dispersion: There are two types of measure of dispersion

i) Absolute Measure of Dispersion
ii) Relative Measure of Dispersion

Absolute Measure of Dispersion: "**An absolute measure of dispersion measures the variability in terms of the same units of the data**" e.g. if the units of the data are Rs, meters, kg, etc. The units of the measures of dispersion will also be Rs, meters, kg, etc. The common absolute measures of dispersion are:

- Range
- Quartile Deviation or Semi Inter-Quartile Range
- Average Deviation or Mean Deviation
- Standard Deviation

➢ **Relative Measure of Dispersion:** **"A relative measure of dispersion compares the variability of two or more data that are independent of the units of measurement"**

➢ In other word **"A relative measure of dispersion, expresses the absolute measure of dispersion relative to the relevant average and multiplied by 100 many times" i.e.**

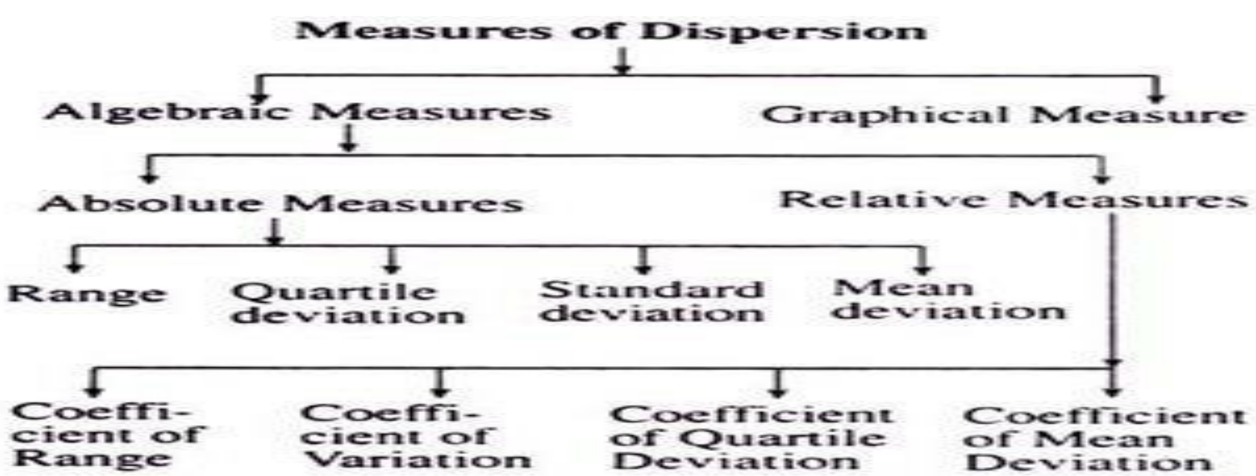
$$Relative\ Dispersion = \frac{Absolute\ Dispersion}{Average}$$

$$Relative\ Dispersion = \frac{Absolute\ Dispersion}{Average} \times 100$$

This is a pure number and independent of the units in which the data has been expressed. It is used for the purpose to compare the dispersion of a data with the dispersion of another data.

Which measures of Dispersion to choose?

Absolute Measure of Dispersion	**Relative Measure of Dispersion**
When dealing with data, if ones' objective is "only to determine" the variation of single set of variable/Information: s/he can/will choose to use Absolute measure of dispersion.	When dealing with data, if ones' objective is "to determine and compare" the variations of multiple set of variables/information having expressed in same/different unit(s): s/he can/will choose to use Relative measure of dispersion.



BRAC UNIVERSITY

Inspiring Excellence

Graphical Measures of dispersion are-

1) Lorenz curve
2) Phillips curve
3) Range Chart
4) Box and Whisker plot etc.

- **Range:** "**The difference between the largest and the smallest value in a set of data is called range"**

OR

"In continuous grouped data the difference between the upper class boundary of the highest class and lower class boundary of the lowest class is called range"

Formula: $R = X_m - X_0$ *OR* *R=Largest value –smallest value*

Where R is the range, X_m is the largest value and X_0 is the smallest value in a dataset.

- **Coefficient of Range or Coefficient of Dispersion:** The coefficient of range or coefficient of dispersion is a relative measure of dispersion and is given by:

$$\text{Coefficient of Range} = \frac{X_m - X_0}{X_m + X_0}$$

Merits	**Demerits**
• Easy to understand and calculate. • It is based only on extreme observations and no detail in formations is required. • It gives us a quick idea of the variability of a set of data	• It is not based on all observation. • Range does not give any indication of the character of the distribution with in the two extreme observations. • Range is subject of fluctuations from sample to sample. • Cannot be computed in case of open-end class.

Numerical example of Range and Coefficient of range

Ex # The marks obtained by 9 students are given below:

x_i	45	32	37	46	39	36	41	48	36

$$R = X_m - X_0$$

$$X_m = 48, \quad X_0 = 32$$

$$R = 48 - 32 \quad \Rightarrow R = 16 \text{ marks}$$

- Coefficient of Range

$$C.o.f = \frac{X_m - X_0}{X_m + X_0}$$

$$C.o.f = \frac{48 - 32}{48 + 32} \quad \Rightarrow C.o.f = \frac{16}{80} \quad \Rightarrow C.o.f = 0.2$$

- **Quartile Deviation or Semi-inter-quartile Range:** "half of the difference between the upper quartile and lower quartile is called the semi-inter quartile range or quartile deviation"i.e.

$$Quartile\ deivation = \frac{Q_3 - Q_1}{2}$$

- **Coefficient of Quartile Deviation:** The coefficient of quartile deviation is a relative measure of dispersion and is given by:

$$Coefficient\ Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Steps of Quartile deviation:

- Quartiles divide the observations in to four equal parts, when observations are arranged in order of magnitudes.
- Median, denoted by Q_2, is the middle most observation and Q_1 & Q_3 are the middle most observations of the lower and upper half respectively.
- Therefore $Q_2 - Q_1$ and $Q_3 - Q_2$ gives us some measure of dispersion.
- The AM of these two measures give us the quartile deviation and is denoted by QD.

$$QD = \frac{(Q_2 - Q_1) + (Q_3 - Q_2)}{2} = \frac{Q_3 - Q_1}{2}$$

BRAC UNIVERSITY
Inspiring Excellence

- **Numerical example of quartile deviation and coefficient of quartile deviation**

Ex # calculate quartile deviation and coefficient of quartile deviation for ungrouped data. The marks obtained by 9 students are given below

x_i	45	32	37	46	39	36	41	48	36

$$Quartile\ deviation = \frac{Q_3 - Q_1}{2}$$

"n=9" is odd then we use odd case formulae

$$Q_1 = Marks\ obtained\ by\ \left[\left(\frac{n}{4}\right)+1\right]^{th}\ student$$ Q_1=36 marks, Q_3=45

$$Quartile\ deviation = \frac{45-36}{2} = 4.5 \text{ marks}$$ (Answer).

- Coefficient of Quartile deviation:

$$Cofficient\ of\ Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$Cofficient\ of\ Q.D = \frac{45-32}{45+32} = 0.11$$ (Answer).

Ex # calculate quartile deviation and coefficient of quartile deviation for continuous grouped data.

Class boundaries	Midpoints (x_i)	Frequency (f_i)	Cumulative frequency $(c.f)$
29.5---39.5	34.5	8	8
39.5---49.5	44.5	87	95
49.5---59.5	54.5	190	285
59.5---69.5	64.5	304	589
69.5---79.5	74.5	211	800
79.5---89.5	84.5	85	885
89.5---99.5	94.5	20	905
		$\sum_{i=1}^{n} f_i$ =905	

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right) \Rightarrow Q_1 = 56.40 \text{ marks}$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - C\right) \Rightarrow Q_3 = 73.76 \text{ marks}$$

Calculate Quartile deviation and coefficient of quartile deviation from above results?

Merits of Quartile deviation:

1. It is simple to understand and easy to calculate.
2. It is not affected by extreme items.
3. It can be calculated for data with open and classes also.

Demerits of Quartile deviation:

1. It is not based on all items. It is based on two positional values Q1 and Q3 and ignores the extreme 50% of the items.
2. It cannot be manipulated algebraically.
3. It is affected by sampling fluctuations.
4. Like range, it does not measure the deviation about any measure of central tendency.

Mean Absolute Deviation or Mean Deviation (Average Deviation): "The arithmetic mean of the absolute deviation from an average (mean, median, mode etc.) is called mean deviation. M.D. is the abbreviation for Mean Deviation. There are three kinds of mean deviations, Viz.,

1. mean deviation or mean deviation about mean, 2. mean deviation about median
3. mean deviation about mode.

	Ungrouped Data	Grouped Data
M.D from Mean	$M.D = \frac{\sum\lvert x_i - \bar{x}\rvert}{n}$	$M.D = \frac{\sum f\lvert x_i - \bar{x}\rvert}{n}$
M.D from Median	$M.D = \frac{\sum\lvert x_i - Med\rvert}{n}$	$M.D = \frac{\sum f\lvert x_i - Med\rvert}{n}$

M.D from Mode   M.D = $\frac{\sum\lvert x_i - Mode\rvert}{n}$ M.D = $\frac{\sum f\lvert x_i - Mode\rvert}{n}$

Coefficient of Mean Deviation: The coefficient of mean deviation is a relative measure of dispersion and is given by:

$$Coefficient\ of\ M.D = \frac{M.D(from\ mean)}{Mean}$$

$$Coefficient\ of\ M.D = \frac{M.D(from\ median)}{Median}$$

$$\text{Coefficient of M.D} = \frac{MD(from\ mode)}{mode}$$

<u>Uses of Mean Deviation:</u>

Mean deviation provides an opportunity to calculate deviation, absolute deviation, total deviation and average of the deviations. Standard deviation is the most important absolute measure of dispersion. knowledge of the principle of mean deviation facilities understanding the concept of standard deviation. Standard deviation is a part of almost all the theories of Statistics, Viz., skewness, kurtosis, correlation, regression, sampling, estimation, inference, S.Q.C., etc. Mean deviation is preferred when a particular discussion is not carried to other spheres. It is found to be much useful in forecasting business cycles and a few other statistical activities connected with business, economic and sociology.

Merits of M.D:

1. Mean deviations are rigidly defined.
2. They are based on all the items.
3. They are affected less by extreme items than standard deviation. Among the three mean deviations, mean deviation about median is the least.
4. They are simple to understand and not difficult to calculate.
5. They do not vary much from sample to sample.
6. They provide choice. among the three mean deviations, the one that is suitable to a particular situation can be used.
7. Formation of different distributions can be compared on the basis of a mean deviation.

Demerits of M.D:

1. Omission of negative sign of deviations makes them non-algebraic. It is pointed out as a great drawback.
2. They could not be manipulated. Combined mean deviation could not be found.
3. It is not widely used in business or economics.

$$\text{Mean Deviation from mean of raw data} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{N}$$

$$\text{Mean deviation from mean of grouped data} = \frac{\sum_{i=1}^{n} [f_i |x_i - \bar{x}|]}{N}$$

where $N = \sum_{i=1}^{n} f_i$, $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} (f_i x_i)$

The following steps are employed to calculate the mean deviation from mean.

Step 1 : Make a column of deviation from the mean, namely $x_i - \bar{x}$ (In case of grouped data take x_i as the mid value of the class.)

Step 2 : Take absolute value of each deviation and write in the column headed $|x_i - \bar{x}|$.
For calculating the mean deviation from the mean of raw data use

$$\text{Mean deviation of Mean} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{N}$$

For grouped data proceed to step 3.

Step 3 : Multiply each entry in step 2 by the corresponding frequency. We obtain $f_i(x_i - \bar{x})$ and write in the column headed $f_i|x_i - \bar{x}|$.

Step 4 : Find the sum of the column in step 3. We obtain $\sum_{i=1}^{n}[f_i|x_i - \bar{x}|]$

Step 5 : Divide the sum obtained in step 4 by N.

Now let us take few examples to explain the above steps.

Example 17.1 Find the mean deviation from the mean of the following data :

Size of items x_i	4	6	8	10	12	14	16
Frequency f_i	2	5	5	3	2	1	4

Mean is 10

Solution :

x_i	f_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
4	2	–5.7	5.7	11.4
6	4	–3.7	3.7	14.8
8	5	–1.7	1.7	8.5
10	3	0.3	0.3	0.9
12	2	2.3	2.3	4.6
14	1	4.3	4.3	4.3
16	4	6.3	6.3	25.2
	21			69.7

Mean deviation from mean $= \frac{\sum[f_i|x_i - \bar{x}|]}{21} = \frac{69.7}{21} = 3.319$

Example 17.2 Calculate the mean deviation from mean of the following distribution :

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of Students	5	8	15	16	6

Mean is 27 marks

Solution :

Marks	Class Marks x_i	f_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
0 – 10	5	5	–22	22	110
10 – 20	15	8	–12	12	96
20 – 30	25	15	–2	2	30
30 – 40	35	16	8	8	128
40 – 50	45	6	18	18	108
Total		50			472

$$\text{Mean deviation from Mean} = \frac{\sum[f_i\,|x_i - \bar{x}|]}{N} = \frac{472}{50}\ \text{Marks} = 9.44\ \text{Marks}$$

Variance: Variance provides an average measure of squared difference between each observation and arithmetic mean. In other words, the variance shows, on an average, how close the values of a variable are to the arithmetic mean. If $X_1, X_2, X_3, \dots\dots\dots, X_N$ are N values of a population of size N, then the population variance, commonly designated as $\boldsymbol{\sigma}^2$, is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}, \text{ where } \mu = \text{Population mean}$$

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{n-1}, \text{ where } \bar{x} = \text{Sample mean}$$

Test Yourself

A population of 10 students got the marks in the examination as given in the table below. Find the variance and Standard Deviation of the given data.

13 15 14 16 2 8 9 23 28 12

Answer:

Step-1: First find the AM of the Population, μ= ??

Step-2: Complete the table.

Step-3: Here, population, N=10

Step-4: Compute variance, $\boldsymbol{\sigma}^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ = ?

Step-5: Compute SD, $\boldsymbol{\sigma}$ = ?

X_i	$X_i - \mu$	$(X_i - \mu)^2$
13		
15		
14		
16		
2		
8		
9		
23		
28		
12		
		$\sum_{i=1}^{N=10}(X_i - \mu)^2$ = ?

Standard Deviation:

Standard deviation, usually denoted by the letter σ (small sigma) of the Greek alphabet was first suggested by Karl Pearson as a measure of dispersion in 1893. It is defined as the positive square root of the arithmetic mean of the squares of the deviations of the given observations from their arithmetic mean. Thus if $X_1, X_2, \ldots, X_n$ is a set of n observations then its standard deviation is given by :

$$\sigma = +\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

The following steps are employed to calculate the variance and hence the standard deviation of raw data. The mean is assumed to have been calculated already.

Step 1 : Make a column of deviations from the mean, namely, $x_i - \bar{x}$.

Step 2 (check) : Sum of deviations from mean must be zero, i.e., $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Step 3: Square each deviation and write in the column headed $(x_i - \bar{x})^2$.

Step 4 : Find the sum of the column in step 3.

Step 5 : Divide the sum obtained in step 4 by the number of observations. We obtain σ^2.

Step 6 : Take the positive square root of σ^2. We obtain σ (Standard deviation).

Example 17.9 The daily sale of sugar in a certain grocery shop is given below :

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
75 kg	120 kg	12 kg	50 kg	70.5 kg	140.5 kg

The average daily sale is 78 Kg. Calculate the variance and the standard deviation of the above data.

Solution : $\bar{x} = 78$ kg (Given)

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
75	– 3	9
120	42	1764
12	– 66	4356
50	– 28	784
70.5	– 7.5	56.25
140.5	62.5	3906.25
	0	10875.50

Thus $\sigma^2 = \frac{\sum_{i=1}(x_i - \bar{x})^2}{n} = \frac{10875.50}{6} = 1812.58$ (approx.)

and $\sigma = 42.57$ (approx.)

Example 17.10 The marks of 10 students of section A in a test in English are given below :

7 10 12 13 15 20 21 28 29 35

Determine the variance and the standard deviation.

Solution : Here $\bar{x} = \frac{\sum x_i}{10} = \frac{190}{10} = 19$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
7	– 12	144
10	– 9	81
12	– 7	49
13	– 6	36
15	– 4	16
20	+1	1
21	+2	4
28	+9	81
29	+10	100
35	+16	256
	0	768

Thus $\sigma^2 = \frac{768}{10} = 76.8$ and $\sigma = +\sqrt{76.8} = 8.76$ (approx)

Test Yourself

A population of 40 students got marks in the examination as given in the table below. Find the variance and Standard Deviation of the given data.

X_i	15	20	25	30	35
f_i	6	8	15	7	4

Answer:

Step-1: First find the AM of the Population, μ= ??

Step-2: Complete the table.

Step-3: Here, population, N=40

Step-4: Compute variance, $\sigma^2 = \frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{N}$ = ? 302.125

Step-5: Compute SD, σ = ? 17.38

X_i	f_i	$X_i - \mu$	$f_i(X_i - \mu)^2$
15	6	7	294
20	8	12	1152
25	15	17	4335
30	7	22	3388
35	4	27	2916
	$\sum_{i=1}^{k=5} f_i$= N = ? 40		$\sum_{i=1}^{k=5} f_i (X_i - \mu)^2$ = ? 12085

Uses of Standard deviation: Standard deviation is the best absolute measure of dispersion. It is a part of many statistical concepts such as Skewness, Kurtosis, Correlation, Regression, Estimation, sampling, tests of Significance and Statistical Quality Control. Not only in statistics but also in Biology, education, Psychology and other disciplines standard deviation is of immense use.

Merits of Standard Deviation:

1. It is rigidly defined;
2. It utilizes all the observations;
3. Squaring is a better technique to get rid of negative deviations; and
4. It is the most popular measure of dispersion.
5. It is calculated on the basis of the magnitudes of all the items.
6. It could be manipulated further. The combined S.D. can be calculated.
7. Mistakes in its calculation can be corrected. The entire calculation need not be redone.
8. Coefficient of variation is based on S.D. It is the best and most widely used relative measure of dispersion.
9. It is free from sampling fluctuations. This property of sampling stability has brought it an indispensable place in tests of significance.
10. It reduces the complexity in the approach of normal distribution by providing standard normal variables.
11. It is the most important absolute measure of dispersion. It is used in all the areas of statistics. It is widely used in other disciplines such as Psychology, Education and Biology as well.
12. Scientific calculators show the standard deviation of any series.
13. Different forms of the formula are available.

Demerits of Standard Deviation:

1. In cases where mean is not a suitable average, standard deviation may not be the appropriate measure of dispersion like when open end classes are present. In such cases quartile deviation may be used;
2. It is not unit free; and
3. Although it is easy to understand but calculation may require a calculator or a computer.
4. Compared with other absolute measures of dispersion, it is difficult to calculate.
5. It is not simple to understand.
6. It gives more weight age to the items away from the mean than those near the mean as the deviations are squared.

Coefficient of Variation: Coefficient of variation is the most widely used relative measure of dispersion. It is based on the best absolute measure of dispersion and the best measure of central tendency. It is a percentage. While comparing two or more groups, the group which has less coefficient of variation is less variable or more consistent or more stable or more uniform or more homogeneous. Coefficient of Variation is denoted by the C.V.

$$\text{C.V.} = \frac{\sigma}{\overline{x}} \times 100, \overline{x} \neq 0$$

where σ and $\overline{x}$ are standard deviation and mean of the data. The coefficients of variation are compared to compare the variability of two series. The series with greater C.V. is said to be more variable than the other. The series having less C.V. is said to be more consistent than the other.

For series with same means, we can have

$$\text{C.V. (1st distribution)} = \frac{\sigma_1}{\overline{x}} \times 100 \qquad ...(1)$$

$$\text{C.V. (2nd distribution)} = \frac{\sigma_2}{\overline{x}} \times 100 \qquad ...(2)$$

where σ_1, σ_2 are standard deviation of the Ist and 2nd distribution respectively, $\overline{x}$ is the equal mean of the distributions.

From (1) and (2), we can conclude that two C.V.'s can be compared on the basis of the values of σ_1 and σ_2 only.

Example 17.21 The standard deviation of two distributions are 21 and 14 and their equal mean is 35. Which of the distributions is more variable?

Solution : Let

σ_1 = Standard dev. of 1st series = 21

σ_2 = Standard dev. of 2nd series = 14

$\overline{x}$ = 35

$$\text{C.V. (Series I)} = \frac{\sigma_1}{\overline{x}} \times 100 = \frac{21}{35} \times 100 = 60$$

$$\text{C.V. (Series II)} = \frac{\sigma_1}{\overline{x}} \times 100 = \frac{14}{35} \times 100 = 40$$

C.V. of series I > C.V. of series II

$\Rightarrow$ Series with S.D = 21 is more variable.

Example 17.22 Monthly wages paid to workers in two factories A and B and other data are given below :

	Factory A	Factory B
Mean of monthly wages	` 15550	` 15550
Variance of the distribution of wages	100	121

Which factory A or B shows greater variablility in individual wages?

Solution : Given

$$\sigma_A = \sqrt{\text{variance}} = \sqrt{100} = 10$$

$$\sigma_B = \sqrt{\text{variance}} = \sqrt{121} = 11$$

$$\bar{x} = \text{`}\ 15550$$

Now,

$$\text{C.V. (A)} = \frac{\sigma_A}{\bar{x}} \times 100 = \frac{10}{15550} \times 100$$

$$= 0.064$$

$$\text{C.V.(B)} = \frac{\sigma_B}{\bar{x}} \times 100 = \frac{11}{15550} \times 100 = 0.07$$

Clearly C.V. (B) > C.V.(A)

$\therefore$ Factory B has greater variability in the individual wages.

Assignment 006:

8. To check the quality of two brands of lightbulbs, their life in burning hours was estimated as under for 100 bulbs of each brand.

Life (in hrs)	*No. of bulbs Brand A*	*Brand B*
0–50	15	2
50–100	20	8
100–150	18	60
150–200	25	25
200–250	22	5
	100	*100*

(i) Which brand gives higher life?
(ii) Which brand is more dependable?

Assignment 006:

Course Number	2	3	4	5	6
Sample 1	2	5	10	12	1
Sample 2	1	6	8	13	2

For which sample of students, the relative variability of course numbers is higher?