Introduction to Dispersion

BRAC UNIVERSITY
Inspiring Excellence

What is Dispersion?

- Measures of average (such as the median and mean) represent the typical value for a dataset. But within the dataset, the actual values usually differ from one another and from the average value itself.
- The extent to which the central value are good representatives of the values in the original dataset depends upon **the variability or dispersion** in the original data.
- Dispersion is the **spread or scatter** of item values from a measure of central tendency. Dispersion is usually measured as an **average of deviations** about some central value.
- Dispersion thus is a type of average and is sometimes called a second order average. Datasets are said to have **high dispersion or variation** when they contain values considerably higher and lower than the central value.

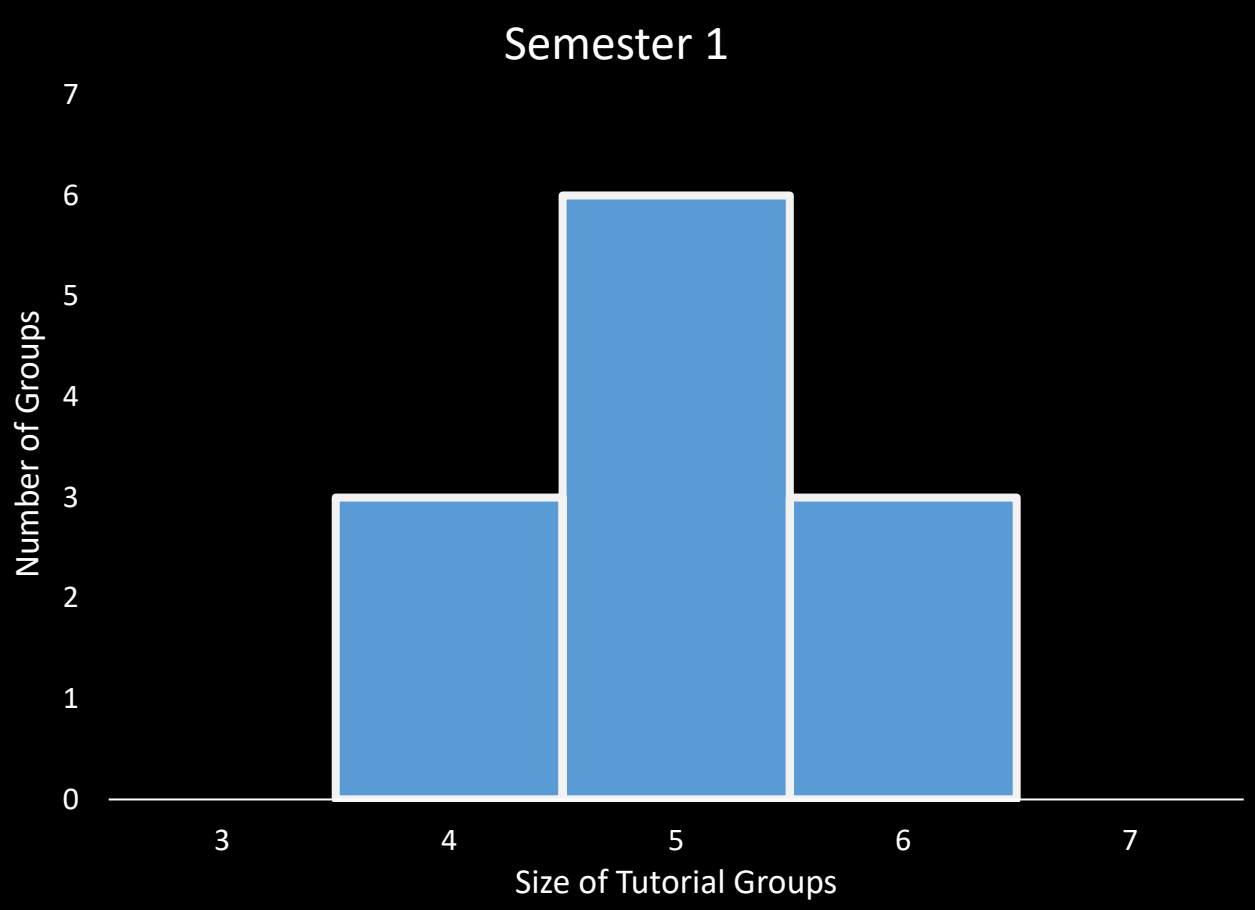
Example 1

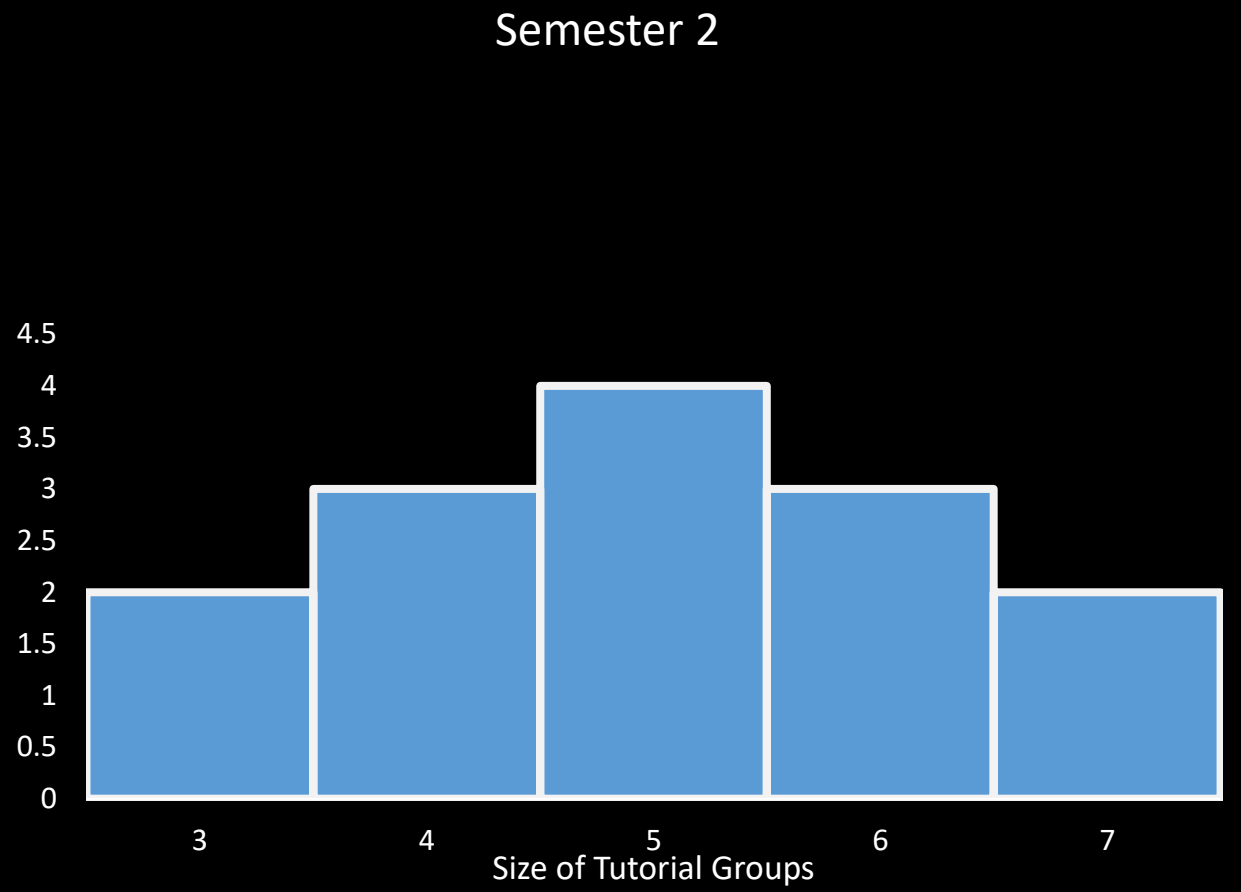
- let us consider two groups of students with score in a particular examination as shown in the table.

Group 1	49	50	50	51
Group 2	0	0	100	100

- The AM for each group is 50.
- It is clear from the data that the first group consists of near average intelligent student and the 2nd group is made up of very bright and very dull students.
- It is evident that the distributions of both groups have the same AM.
- But they differ in variation from $\overline{X}$; such variation is usually measured by the measure of dispersion.

Example 2





BRAC UNIVERSITY
Inspiring Excellence

Example 2

- In the two charts, the number of different sized tutorial groups in semester 1 and semester 2 are presented.
- In both semesters the mean and median tutorial group size is 5 students, however the groups in semester 2 show more dispersion (or variability in size) than those in semester 1.

Characteristics of a good measure of dispersion

- Dispersion within a dataset can be measured or described in several ways including the **range, inter-quartile range, variance, and standard deviation.**
- 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.

Purpose of measure of dispersion

Measure of dispersion is important for the following purpose.

i. To determine the reliability of an average.

ii. To compare the variability.

iii. To compare two or more series with regard to their variability.

iv. To facilitate the use of other statistical measures.

v. It is one of the most important quantities used to characterize a frequency distribution.

Types of measure of dispersion

- Measure of dispersion or variation may be either absolute or relative.
- **Absolute measure** of variation is expressed in the same statistical unit in which the original data are given such as takas, kilograms, tones, etc. and may be used to compare the variation in two distributions, provided the variables are expressed in the same units and of same average size.
- On the other hand, often it is necessary to compare the distribution in two or more different frequency distributions having variables expressed in different units.
- In such a case dispersion is calculated by *dividing the absolute measure of dispersion by a measure of central tendency* – which generates pure number that are independent of the unit of measurement. The resultant numerical value is a **relative measure** of dispersion.

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

Different types of Absolute and Relative measure of dispersion

Absolute Measure of Dispersion	Relative Measure of Dispersion
1. Range 2. Quartile deviation 3. Variance and Standard deviation	1. Coefficient of range 2. Coefficient of quartile deviation 3. Coefficient of variation and standard deviation

Range

- The range of a set of data values is the difference between the highest and the lowest values in the set.
- If X_l & X_s are the smallest and the largest values respectively in a set then the range "R" is defined as $\mathrm{R} = X_l - X_s$.
- For group data the range is taken either as the difference between the lower boundary of the first class and the upper boundary of the last class or as the difference between the highest and the lowest mid-values

Coefficient of Range

- The coefficient of dispersion corresponding to range called coefficient of range

$$\text{Coefficient of Range} = \frac{X_l - X_s}{X_l + X_s};\ \text{Where } X_l = \text{Largest value and } X_s = \text{Smallest Value}$$

Range

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.

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.

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

Coefficient of Quartile Deviation

- The coefficient of variation corresponding to quartile deviation is called the coefficient of quartile deviation and is defined as

$$\text{Coefficient of } QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Quartile Deviation

Merits	Demerits
• It is superior to range as a measure of dispersion. • It is applicable in Open-end class. • Easy to understand and compute. • Not affected by extreme values.	• It ignores 50% of items that is the first 25% and last 25% of observations. • Very much affected by sampling fluctuations. • Not suited for further algebraic treatment.

Test Yourself

The following data represents Annual wages of two Factories X and Y. for the given information

I. Determine range and coefficient of range. (in ‘000 Tk)

II. Determine the quartile deviation and coefficient of Co-efficient of quartile deviation.

Test Yourself

Table 1: Annual wages of Factory X workers (in ‘000 Tk)

91	70	74	79	86	93
60	71	76	79	87	96
112	72	127	79	87	62
68	72	77	79	90	76
69	73	77	85	48	157

Table 2: Annual wages of Factory Y workers (in ‘000 Tk)

97	78	85	92	97	105
72	79	85	92	97	107
112	79	87	92	97	72
113	80	90	96	68	75
78	82	90	97	100	

Hints:

1. Find the lowest and highest value for each table.
2. Calculate the quartiles (Q_1, Q_2, Q_3) for each table.

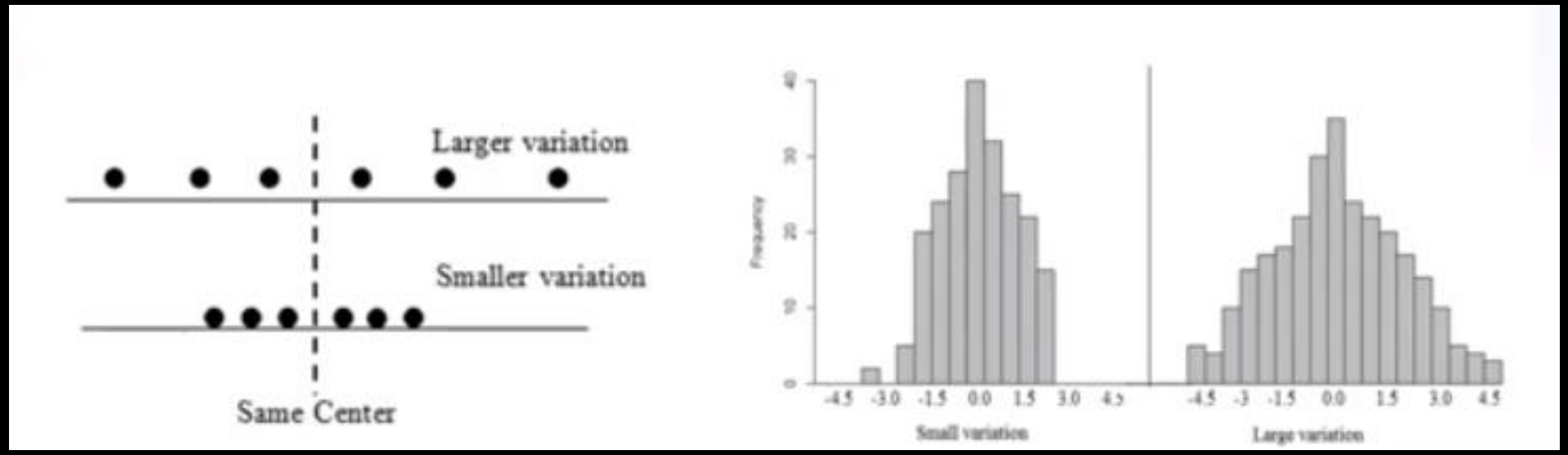
BRAC UNIVERSITY
Inspiring Excellence

Variance and Standard Deviation

BRAC UNIVERSITY
Inspiring Excellence

Variance

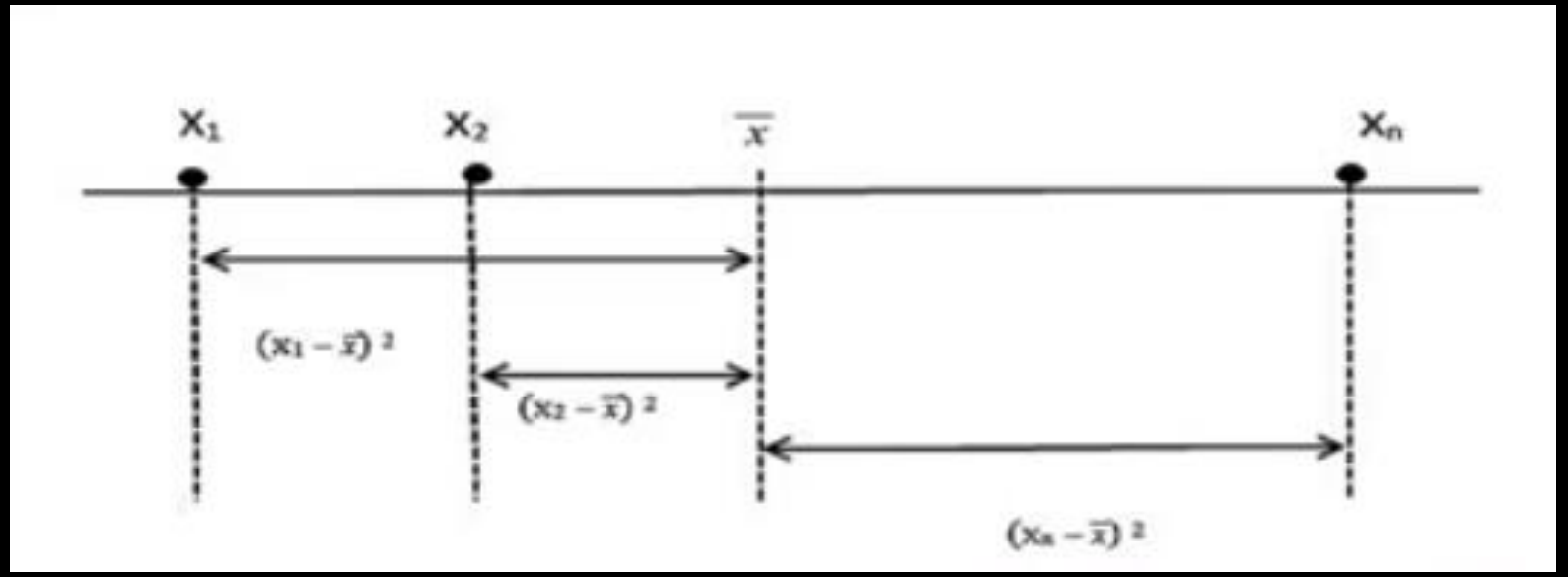
- Variance provides an average measure of squared difference between each observation and arithmetic mean.
- In other words, the variance shows, on an average, how close the values of a variable are to the arithmetic mean.



Calculating Variance

- If $x_1, x_2, x_3, \ldots\ldots\ldots, x_N$ are sample values and $\bar{x}$ is the sample mean, then the deviation of the value x_i from the sample mean $\bar{x}$ is $(x_i - \bar{x})$ and the squared deviation is $(x_i - \bar{x})^2$.

- The sum of squared deviations is $\sum_{i=1}^{n}(x_i - \bar{x})^2$. The following graph shows the squared deviations of the values from their mean.



Standard Deviation

- The variance represents squared units, and therefore is not appropriate measure of dispersion when we wish to express the concept of dispersion in terms of the original unit.
- The Standard deviation is another measure of dispersion. The standard deviation is the positive square root of the variance and is expressed in the original unit of the data.'
- Standard Deviation of variable X, $SD(X) = \sqrt{VAR(X)}$

Population: Ungrouped Data

- 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

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

  - For the same population, Standard Deviation (SD) of the population, commonly designated as $\boldsymbol{\sigma}$, is defined as

$$\boldsymbol{\sigma} = \sqrt{Variance} = \sqrt{\boldsymbol{\sigma^2}} = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

BRAC UNIVERSITY
Inspiring Excellence

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		
		$\sum_{i=1}^{N=10}(X_i - \mu)^2 = ?$

BRAC UNIVERSITY
Inspiring Excellence

Population: Grouped Data

- In case of grouped data, if $X_1, X_2, X_3, \dots\dots\dots, X_k$ are values that occur with frequencies $f_1, f_2, f_3, \dots\dots\dots, f_k$ respectively in a population of size N, then the population variance, $\boldsymbol{\sigma}^2$, is defined as

$$\boldsymbol{\sigma}^2 = \frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{N}$$

- For the same population, Standard Deviation (SD) of the population, $\boldsymbol{\sigma}$, is defined as

$$\boldsymbol{\sigma} = \sqrt{Variance} = \sqrt{\boldsymbol{\sigma}^2} = \sqrt{\frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{N}}$$

BRAC UNIVERSITY
Inspiring Excellence

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, $\boldsymbol{\sigma}^2 =$

$\frac{\sum_{i=1}^{k} f_i(X_i - \mu)^2}{N} = ?$

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

X_i	f_i	$X_i - \mu$	$f_i(X_i - \mu)^2$
15	6		
20	8		
25	15		
30	7		
35	4		
	$\sum_{i=1}^{k=5} f_i = \mathrm{N} = ?$		$\sum_{i=1}^{k=5} f_i(X_i - \mu)^2 = ?$

BRAC UNIVERSITY
Inspiring Excellence

Sample: Ungrouped Data

- If $X_1, X_2, X_3, \ldots\ldots\ldots, X_n$ are n values of a sample of size n, then the sample variance, commonly designated as $\boldsymbol{s^2}$, is defined as

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

  - For the same population, Standard Deviation (SD) of the population, commonly designated as $\boldsymbol{s}$, is defined as

$$\boldsymbol{s} = \sqrt{Variance} = \sqrt{\boldsymbol{s^2}} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n-1}}$$

BRAC UNIVERSITY
Inspiring Excellence

Test Yourself

A sample of 10 students' marks in the examination is 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 Sample, $\bar{x}$= ??

Step-2: Complete the table.

Step-3: Here, sample size, n=10

Step-4: Compute variance, $\boldsymbol{s}^2=$ $\frac{\sum_{i=1}^{n}(X_i-\bar{x})^2}{n-1}$ = ?

Step-5: Compute SD, s = ?

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

BRAC UNIVERSITY
Inspiring Excellence

Sample: Grouped Data

- In case of grouped data, if $X_1, X_2, X_3, \dots\dots\dots, X_k$ are values that occur with frequencies $f_1, f_2, f_3, \dots\dots\dots, f_k$ respectively in a sample of size n, then the sample variance, $\boldsymbol{s^2}$, is defined as

$$\boldsymbol{s^2} = \frac{\sum_{i=1}^{k} f_i (X_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1} = \frac{\sum_{i=1}^{k} f_i (X_i - \bar{x})^2}{n-1}$$

- For the same population, Standard Deviation (SD) of the population, s, is defined as

$$\boldsymbol{s} = \sqrt{Variance} = \sqrt{\boldsymbol{s^2}} = \sqrt{\frac{\sum_{i=1}^{k} f_i (X_i - \bar{x})^2}{n-1}}$$

BRAC UNIVERSITY
Inspiring Excellence

Test Yourself

A sample of 40 students' marks in the examination is 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 Sample, $\bar{x}$= ??

Step-2: Complete the table.

Step-3: Here, sample size, n=40

Step-4: Compute variance, $\boldsymbol{s}^2$= $\frac{\sum_{i=1}^{k} f_i(X_i-\bar{x})^2}{n-1}$= ?

Step-5: Compute SD, s = ?

X_i	f_i	$X_i - \bar{x}$	$f_i(X_i - \bar{x})^2$
15	6		
20	8		
25	15		
30	7		
35	4		
	$\sum_{i=1}^{k=5} f_i$= n = ?		$\sum_{i=1}^{k=5} f_i(X_i - \bar{x})^2$= ?

BRAC UNIVERSITY
Inspiring Excellence

Population or Sample?

- As we can see, the formulas for variance and SD differ between population and sample.
- For population, it is a parameter, whereas for sample, it is a statistic.
- Unless clearly mentioned, datasets are samples taken from a large population.
- So the formula for populations should be used:
  - If the question directly mentions that the data is for the whole population or
  - If the question dictates that the data was taken for all members of population e.g. all students of a class, and then ask for variance/SD for that population.
- Otherwise, always use the formula for samples, specially when nothing is mentioned about sample/population.

Why $n-1$, not n?

- We see that both formulas of samples' variance/SD (grouped/ungrouped) has $n-1$ as a denominator, instead of n.
- But for population, it is only N, the total population.
- The reason for using $n-1$ is complex and out of scope, but three basic points can be mentioned:
  I. The average of all variance/SD taken from sample should be equal to the variance/SD of the population. If we use n, it is not.
  II. For $n-1$, the sample's variance/SD is closer to population's.
  III. As samples are a finite set from the population, the value of the last data is determined by the value of others. Thus the *degree of freedom* of the set is 1 less than the size i.e. $n-1$.

Test Yourself

An Advertising company is looking for a group of extras to shoot a sequence for a movie. The ages of the first 20 candidates to be interviewed are

50	56	44	49	52	57	56	57	56	59
54	55	61	60	51	59	62	52	54	49

The director of the movie wants men whose ages are tightly grouped around 55 years. Being a statistics buff of sorts, the director suggests that a standard deviation of 3 years would be acceptable. Does this group of extras qualify?

Hints: Calculate AM, Variance, and SD. Then compare the SD with 3 as given in the problem.

Coefficient of Variation (CV)

- The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean:

$$\text{Population CV, } c_v = \frac{\sigma}{\mu} * 100 = value\ in\ x\%$$

$$\text{Sample CV, } c_v = \frac{s}{\bar{x}} * 100 = value\ in\ x\%$$

- It shows the extent of variability in relation to the mean of the population.
- The coefficient of variation should be computed only for data measured on a ratio scale.
- For comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.

Comparing Coefficients

- C.V. is free of unit (unit-less) and the SD is divided by the corresponding AM to make it comparable for different means.
- To compare the variability of two sets of data (i.e. to determine which set is more variable), we need to calculate the AM and SD of both sets.
- Then we can calculate the CVs for both sets.
- The data set with the larger value of CV has larger variation which is expressed in percentage.
- For example, the relative variability of the data set 1 will be larger than data set 2 if $CV_1 > CV_2$ and vice versa.

Test Yourself

In a University, students can take any number of courses per semester. For two samples of 30 students each. the data of how many courses one takes is given below:

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?

Hints: Calculate AM, Variance, and SD for both samples. Then compare the CVs.

BRAC UNIVERSITY
Inspiring Excellence

Combined Standard Deviation

- The combined standard deviation of two sets of data containing n_1 and n_2 observations with means μ_1 and μ_2 and standard deviations $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ respectively is given by

$$\boldsymbol{\sigma_{12}} = \sqrt{\frac{n_1(\boldsymbol{\sigma}_1^2+d_1^2)+n_2(\boldsymbol{\sigma}_2^2+d_2^2)}{n_1+n_2}}$$

Where, $\boldsymbol{\sigma_{12}}$ = combined Standard Deviation

$d_1 = \mu_{12} - \mu_1$

$d_2 = \mu_{12} - \mu_2$

And $\mu_{12} = \frac{\boldsymbol{n_1}\mu_1 + \boldsymbol{n_2}\mu_2}{n_1+n_2}$

- This formula of combined standard deviation of two sets of data can be extended to compute the standard deviation of more than two sets of data on the same lines.

BRAC UNIVERSITY
Inspiring Excellence

Test Yourself

From the analysis of monthly wages paid to employees in two service organizations X and Y, the following results were obtained:

	Organization X	Organization Y
Number of wage-earners	550	650
Average monthly wages	5000	4500
Variance of the distribution of wages	900	1600

a. Which organization pays a larger amount as monthly wages?

b. Determine the combined variance of all the employees taken together?

Test Yourself

For a group of 50 male workers, the mean and standard deviation of their monthly wages are tk. 6300 and tk. 600 respectively. For a group of 40 female workers, these are tk. 5400 and tk. 600 respectively. Find the standard deviation of monthly wages for the combined group of workers.

Application of Standard Deviation

- In stock charts, Standard Deviation (SD) is a measure of volatility. Chartists can use SD to measure expected risk and determine the significance of certain price movements.
- Bollinger Bands show the upper and lower limits of 'normal' price movements based on SD of prices.
- In many fields of Science and Business studies, SD is significant. From the variance and SD we can understand the fitness of a statistical model when we deal with the dependencies of data.

Variance

Merits	Demerits
• Rigidly defined. • Based upon all observation. • Easy to understand • Less affected by sampling fluctuations. • Suitable for further algebraic treatment.	• Difficult to calculate. • Affected by extreme values. • Difficult to calculate for open-end class.