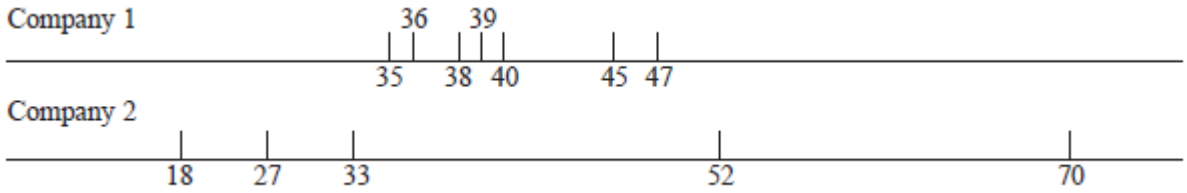
Measurement of Dispersion for Ungrouped Data

The measures of central tendency, such as the mean, median, and mode, do not reveal the whole picture of the distribution of a data set. Two data sets with the same mean may have completely different spreads. The variation among the values of observations for one data set may be much larger or smaller than for the other data set. (Note that the words dispersion, spread, and variation have the same meaning.) Consider the following two data sets on the ages (in years) of all workers working for each of two small companies.

Company 1:	47	38	35	40	36	45	39
Company 2:		70	33	18	52	27	

The mean age of workers in both these companies is the same, 40 years. If we do not know the ages of individual workers at these two companies and are told only that the mean age of the workers at both companies is the same, we may deduce that the workers at these two companies have a similar age distribution. As we can observe, however, the variation in the workers' ages for each of these two companies is very different. As illustrated in the diagram, the ages of the workers at the second company have a much larger variation than the ages of the workers at the first company.



Thus, the mean, median, or mode by itself is usually not a sufficient measure to reveal the shape of the distribution of a data set. We also need a measure that can provide some information about the variation among data values. The measures that help us learn about the spread of a data set are called the measures of dispersion. The measures of central tendency and dispersion taken together give a better picture of a data set than the measures of central tendency alone.

This section discusses three measures of dispersion: range, variance, and standard deviation.

1- Range

The range is the simplest measure of dispersion to calculate. It is obtained by taking the difference between the largest and the smallest values in a data set.

Finding the Range for Ungrouped Data

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

EXAMPLE [1]

Table(1) gives the total areas in square miles of the four western South-Central states of the United States

State	Total Area (square miles)
Arkansas	53,182
Louisiana	49,651
Oklahoma	69,903
Texas	267,277

Find the range for this data set.

Solution: The maximum total area for a state in this data set is 267,277 square miles, and the smallest area is 49,651 square miles. Therefore,

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$
$$= 267{,}277 - 49{,}651 = \mathbf{217{,}626 \text{ square miles}}$$

Thus, the total areas of these four states are spread over a range of 217,626 square miles.

The range, like the mean, has the disadvantage of being influenced by outliers. In Example [1], if the state of Texas with a total area of 267,277 square miles is dropped, the range decreases from 217,626 square miles to 20,252 square miles. Consequently, the range is not a good measure of dispersion to use for a data set that contains outliers.

Another disadvantage of using the range as a measure of dispersion is that its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range. Thus, the range is not a very satisfactory measure of dispersion.

2- Variance and Standard Deviation

The **standard deviation** is the most-used measure of dispersion. The value of the standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the

standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.

The *standard deviation is obtained by taking the positive square root of the* **variance**. The variance calculated for population data is denoted by $\boldsymbol{\sigma^2}$ (read as *sigma squared*)^2 and the variance calculated for sample data is denoted by $\boldsymbol{s^2}$. Consequently, the standard deviation calculated for population data is denoted by $\boldsymbol{\sigma}$ and the standard deviation calculated for sample data is denoted by $\boldsymbol{s}$. Following are what we will call the ***basic formulas*** that are used to calculate the variance

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

The quantity $x - \mu$ or $x - \bar{x}$ in the above formulas is called the *deviation* of the x value from the mean. The sum of the deviations of the x values from the mean is always zero; that is, $\Sigma(x-\mu) = 0$ and $\Sigma(x-\bar{x}) = 0$.

For example, suppose the midterm scores of a sample of four students are 82, 95, 67, and 92, respectively. Then, the mean score for these four students is

$$\bar{x} = \frac{82 + 95 + 67 + 92}{4} = 84$$

The deviations of the four scores from the mean are calculated in Table (2) below. As we can observe from the table, the sum of the deviations of the *x* values from the mean is zero; that is,

$$\Sigma(x - \bar{x}) = 0.$$

For this reason we square the deviations to calculate the variance and standard deviation.

x	$x - \bar{x}$
82	$82 - 84 = -2$
95	$95 - 84 = +11$
67	$67 - 84 = -17$
92	$92 - 84 = +8$
	$\Sigma(x - \bar{x}) = 0$

From the computational point of view, it is easier and more efficient to use *short-cut formulas* to calculate the variance and standard deviation. By using the short-cut formulas, we reduce the computation time and round-off errors. The short-cut formulas for calculating the variance and standard deviation are given next.

Short-Cut Formulas for the Variance and Standard Deviation for Ungrouped Data

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

Note that the denominator in the formula for the population variance is N, but that in the formula for the sample variance it is n - 1.

EXAMPLE 3–12

The following table(3) gives the 2008 market values (rounded to billions of dollars) of five international companies

Company	Market Value (billions of dollars)
PepsiCo	75
Google	107
PetroChina	271
Johnson & Johnson	138
Intel	71

Find the variance and standard deviation for these data.

Solution Let x denote the 2008 market value (in billions of dollars) of a company. The values of $\sum x$ and $\sum x^2$ are calculated in Table(4)

x	x^2
75	5625
107	11,449
271	73,441
138	19,044
71	5041
$\Sigma x = 662$	$\Sigma x^2 = 114{,}600$

Calculation of the variance involves the following four steps

Step 1. *Calculate* $\sum x$

The sum of the values in the first column of Table (4) gives the value of $\sum x$ which is 662.

Step 2. *Find* $\sum x^2$

The value of $\sum x^2$ is obtained by squaring each value of x and then adding the squared values.

The results of this step are shown in the second column of Table (4). Notice that $\sum x^2 = 114{,}600$.

Step 3. *Determine the variance.*

Substitute all the values in the variance formula and simplify. Because the given data are on the market values of only five companies, we use the formula for the sample variance.

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{114{,}600 - \frac{(662)^2}{5}}{5-1} = \frac{114{,}600 - 87{,}648.80}{4} = \mathbf{6737.80}$$

Step 4. Obtain the standard deviation.

The standard deviation is obtained by taking the (positive) square root of the variance.

$$s = \sqrt{6737.80} = 82.0841 = \mathbf{\$82.08\ billion}$$

Thus, the standard deviation of the market values of these five companies is $82.08 billion

Two Observations

1- *The values of the variance and the standard deviation are never negative.*
2- *The measurement units of variance are always the square of the measurement units of the original data.*