Measurement of Dispersions

The descriptive measures of another defining characteristic of a data set: how dispersed or spread out the data are. This measures are measures of dispersion, or measures of variation, or measures of variability. The most frequently used measures of variation are:

- Range
- Standard deviation
- Variance
- Percentile, deciles and quartiles

1 Range

The sample range is obtained by computing the difference between the largest value and smallest values in the data set.

Range = Maximum value – Minimum value

The sample range of the variable is quite easy to compute. However, in using the range, many data values are ignored. This is because only the largest and smallest values of the variable are considered. The other observed values are disregarded.

Example 1

The following is a set of recorded of time in minutes of the participants in bike race.
28, 22, 29, 21, 24, 25, 23 and 27.
Find the range of the bike race.

Solution

The maximum value = 29 minutes
The minimum value = 21 minutes

Range = max. – min.
= 29 – 21 minutes
Range = 8 minutes

Limitations of Range

The range is a good way to get a very basic understanding of how spread out our data are. It is easy to calculate as it only requires a basic arithmetic operation. The range is a very crude measurement of the spread of data because it is extremely sensitive to outliers. A single data value can greatly affect the value of the range.

For example, consider the set of data 1, 2, 3, 4, 6, 7, 7, 8.
The maximum value = 8
the minimum value = 1
∴ the range = 8 – 1 = 7

Now consider the same set of data, with the value 100 included.
the new set of data:
1, 2, 3, 4, 6, 7, 7, 8, 100

The maximum value = 100
the minimum value = 1

The range now = 100 – 1 = 99.

The addition of a single extra data point greatly affected the value of the range.

2 Variance and Standard Deviation

Variance and standard deviation are two popular measures of variation. Their formulations are categorized into whether to evaluate from a population or from a sample.

Measures of Variability

1. **Variance:**

Variance is a number that gives a general idea of how the values in a data set are spread out. And the larger the variance the more dispersed the data is. The spread of data tell us how much the individual numbers of the data set differ.

Variance of a set of observations is the average squared deviation of the data points from their mean.

Population variance means that we will collect every member of the population in the data set and sample variance means that the data was extracted from a sample of the population.

2. **Standard deviation:**

The square root of the variance is the standard deviation.

Population Variance (σ^2)

Population variance (σ^2) is the mean square of all deviations from the mean.

The formula for the population variance is:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Where:

σ^2 stands for the population variance

μ stands for the population mean (μ read mu)

x_i stands for a particular value

Σ is the Greek capital sigma and indicates the operation of adding.

N is the total number of values in the population.

Sample Variance (S^2)

Sample variance **(S^2)** is the mean square of all deviations from the sample mean.

The formula for the sample variance is:

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

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

For ungrouped data:

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

or

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

Where:

S^2 stands for the sample variance

$\bar{x}$ stands for the sample mean ($\bar{x}$ read *x* bar)

x_i stands for a particular value

Σ is the Greek capital sigma and indicates the operation of adding.

n is the total number of values in the sample.

Standard deviation

We use standard deviation to measure the spread of a set of data from its mean. If a set of data is widespread, then the deviation of data is quite high. Standard deviation is useful when comparing the spread of two data sets that have approximately the same mean.

1. The population standard deviation = $\sqrt{Population\ variance}$

$$\sigma = \sqrt{\sigma^2}$$

2. The sample standard deviation = $\sqrt{Sample\ variance}$

$$S = \sqrt{S^2}$$

The formulas for the standard deviation are as following.

1. Population standard deviation

The formula for the population standard deviation is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Where:

σ^2 stands for the population variance

σ stands for the population standard deviation

μ stands for the population mean (μ read mu)

x_i stands for a particular value

Σ is the Greek capital sigma and indicates the operation of adding.

N is the total number of values in the population.

2. **Sample standard deviation**

Where:

$\bar{x}$ stands for the sample mean ($\bar{x}$ read *x* bar)

x_i stands for a particular value

Σ is the Greek capital sigma and indicates the operation of adding.

$\sum x$ stands for the sum of all the x_i

$\boldsymbol{n}$ is the total number of values in the sample.

1. Ungrouped data:

The formula for the sample standard deviation is:

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

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

or

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

2. Grouped data:

The formula for the sample standard deviation is:

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

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Or $$S = \sqrt{S^2} = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx_i)^2}{n}}{n-1}}$$

Example 2

The quiz scores on mathematics of a sample group of 8 students were: 7, 6, 4, 6, 10, 8, 11, and 12.
Find the mean, standard deviation, and variance of this group.

Solution

The quiz scores: 7, 6, 4, 6, 10, 8, 11, and 12.

n = 8

mean $(\bar{x}) = \frac{\sum x}{n}$

$\bar{x} = \frac{7+6+4+6+10+8+11+12}{8}$

$= \frac{64}{8}$

mean of the quiz score $(\bar{x})$ = 8

Find standard deviation (s):
Method 1:

Scores (x)	x - $\bar{x}$	$(x - \bar{x})^2$
4	4 – 8 = -4	16
6	6 – 8 = -2	4
6	6 – 8 = -2	4
7	7 – 8 = -1	1
8	8 – 8 = 0	0
10	10 – 8 = 2	4
11	11 – 8 = 3	9
12	12 – 8 = 4	16
		$\sum(x - \bar{x})^2 = 54$

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

$= \sqrt{\frac{54}{8-1}} = \sqrt{7.71}$

Standard deviation (S) = 2.77

Method 2:
We shall use the formula as follows:

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

Scores (x)	x^2
4	16
6	36
6	36
7	49
8	64
10	100
11	121
12	144
$\sum x = 64$	$\sum x^2 = 566$

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

$$= \sqrt{\frac{566 - \frac{(64)^2}{8}}{8-1}}$$

$$= \sqrt{\frac{54}{7}}$$

$$= \sqrt{7.714}$$

Standard deviation (S) = 2.77

Sample Variance (s^2) = $(2.77)^2$

= 7.71

Statistical Dispersion

Dispersion refers to the idea that there is a second number which tell us how "**spread out**" all the measurements are from that central number.

The Standard Deviation and Variance

The **standard deviation** is the "**average**" degree to which scores deviate from the mean. We measure how far all of the measurements are from the mean, square each one, and add them all up. The result is called the **variance**. Take square root of the variance, and we have the standard deviation.

Example 3

The data in the table below shows information about the gas consumption of the 100 sample families. Find the mean and sample standard deviation.

Gas Consumption	Number of family
10-19	1
20-29	0
30-39	1
40-49	4
50-59	7
60-69	16
70-79	19
80-89	20
90-99	17
100-109	11
110-119	3
120-129	1

Solution

The number of family (n) = 100

Sample mean $(\bar{x})$ = $\frac{\sum fx}{n}$

We shall use the formula of grouped data to find the sample standard deviation:

$$\mathbf{S} = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx_i)^2}{n}}{\mathbf{n-1}}}$$

Gas Consumption	Number of family (f)	Class boundary	Midpoint (x_i)	$f_i x_i$	$f_i x_i^2$
10-19	1	9.5-19.5	14.5	14.5	210.2
20-29	0	19.5-29.5	24.5	0	0
30-39	1	29.5-39.5	34.5	34.5	1,190.2
40-49	4	39.5-49.5	44.5	178.0	7,921.0
50-59	7	49.5-59.5	54.5	381.5	20,791.7
60-69	16	59.5-69.5	64.5	1032.0	66,564.0
70-79	19	69.5-79.5	74.5	1415.5	105,454.8
80-89	20	79.5-89.5	84.5	1690.0	142,805.0
90-99	17	89.5-99.5	94.5	1606.5	151,814.3
100-109	11	99.5-109.5	104.5	1149.5	120,122.8
110-119	3	109.5-119.5	114.5	343.5	39,330.7
120-129	1	119.5-129.5	124.5	124.5	15,500.2
	100			**7,970**	**671,705**

The sample mean of gas assumption $(\bar{x})$ $= \frac{\sum fx}{n}$

$= \frac{7,970}{100}$

The sample mean of gas assumption $(\bar{x})$ = 79.70

Find the sample standard deviation:
From the table above:

$\sum fx$ = 7,970

$\sum fx^2$ = 671,705

and n = 100

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

$$= \sqrt{\frac{671{,}705 - \frac{(7{,}970)^2}{100}}{100-1}}$$

$$= \sqrt{\frac{671{,}705 - 635{,}209}{99}}$$

$$= \sqrt{368.646} = 19.2$$

sample standard deviation = 19.2 ▣

Example 4 The salary of employees in Company A and Company B are:

Company A Salary (Baht)	5,000	15,000	25,000	35,000	45,000	55,000
Company B Salary (Baht)	5,000	5,000	5,000	55,000	55,000	55,000

a) Find mean of the salary of employees for Company A and Company B

b) What are the standard deviations of the two sample set of data?

c) Is it meaningful to compare the standard deviations of the two set of data?

Solution

a) Mean of the salary of employees for Company A

$$= \frac{5000+15000+25000+35000+45000+55000}{6}$$

$$= \frac{180,000}{6}$$

$= 30,000$ Baht

Mean of the salary of employees for Company B

$$= \frac{5000+5000+5000+55000+55000+55000}{6}$$

$$= \frac{180,000}{6}$$

$= 30,000$ Baht

b) Find the standard deviations of the two sample set of data

Standard deviation of Company A:

Number	Salary (x)	$x - \bar{x}$	$(x - \bar{x})^2$
1	5,000	-25,000	625,000,000
2	15,000	-15,000	225,000,000
3	25,000	-5,000	25,000,000
4	35,000	5,000	25,000,000
5	45,000	15,000	225,000,000
6	55,000	25,000	625,000,000
	$\sum x$ = 180,000		1,750,000,000

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

$= \sqrt{\dfrac{1{,}750{,}000{,}000}{6-1}}$

$= \sqrt{350{,}000{,}000}$

Standard deviation (S) = 18,708.29

Standard deviation of Company B:

Number	Salary (x)	x - $\bar{x}$	$(x - \bar{x})^2$
1	5,000	-25,000	625,000
2	5,000	-25,000	625,000
3	5,000	-25,000	625,000
4	55,000	25,000	625,000
5	55,000	25,000	625,000
6	55,000	25,000	625,000
	$\sum x$ = 180,000		3,750,000,000

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

$= \sqrt{\dfrac{3{,}750{,}000{,}000}{6-1}}$

$= \sqrt{750{,}000{,}000}$

Standard deviation (S) = 27,386.12

c) Is it meaningful to compare the standard deviations of the two set of data?

The mean salary of employee Company A and mean salary Company of employee B is the same value (30,000 Baht), but the **standard deviation** is different values. **Company A**: a smaller standard deviation (18,708.29), has a narrower spread of data around the mean. However, **Company B** has a higher standard deviation (27,386.12), and the data salaries are far away from the mean. The data is spreaded widely. Therefore, it is meaningful to compare the standard deviations of the two sets of data.

Exercise

1. The weekly salary of a sample group of 60 workers in a factory are shown in the table below.

Salary (Baht)	Frequency
$10,000 < x \leq 12,000$	8
$12,000 < x \leq 14,000$	23
$14,000 < x \leq 16,000$	16
$16,000 < x \leq 18,000$	10
$18,000 < x \leq 20,000$	3

Find the mean and the standard deviation.

2. The average of the sample group of 7 girls are 16, 21, 22, 20, 18, x and $2x$ years old. If the mean is 19, find the value of x and the standard deviation of the ages of the girls.

3. Given that six numbers 6, 15, x, 18, 10 and y have a mean of 9, find the value of $x + y$. if the six numbers have a standard deviation of 6, find the values of x and of y.

4. Two trains A and B, are scheduled to arrive at a station at a certain time. The time in minutes after the scheduled time for each of the 40 days was recorded and the results are as follows:

Time (min.)	Number of days for Train A	Number of days for Train B
$2 < x \leq 5$	3	4
$5 < x \leq 8$	2	3
$8 < x \leq 11$	5	9
$11 < x \leq 14$	11	8
$14 < x \leq 17$	9	7
$17 < x \leq 20$	6	5
$20 < x \leq 23$	1	3
$23 < x \leq 27$	2	1
$27 < x \leq 30$	1	0

a) Calculate the mean and standard deviation for each train.
b) Which train is more consistent in arriving late? Explain briefly.
c) Based on these data, which train is more punctual on the whole? Why?

5. The following table shows the life span, to the nearest hour, for 100 light bulbs from two factories:

Life Spans (hour)	Number of the bulbs Factory A	Number of the bulbs Factory B
600 – 699	2	8
700 – 799	9	10
800 – 899	15	12
900 – 999	22	16
1000 – 1099	28	w
1100 – 1199	19	18
1200 – 1299	5	12
Mean	x	989.5
Standard deviation	y	z

a) Find the values of w, x, y, and z
b) Compare and comment briefly on the life spans of the light bulbs produced by the two factories.