4. Measure of Central Tendency and Dispersion

In the case of quantitative data, the central tendancy and dispersion of the data are measured and analyzed.

- Central tendancy : mean / median
- Dispersion : variance / standard deviation

Data for two quantitative variables are measured using a covarince and a correlation coefficient.

4.1 Measure of Central Tendency – Mean and Median

☞ Think	The data obtained by taking a sample of 5 middle school students and surveying their weight are as follows. (Data 4.1) Weights of five middle school students (kg) 63 60 65 55 77
Explore	1) What kind of graph is used to find a representative value of these data? 2) What would be a representative value for the weight of 5 students?

- The **average** (or **mean**) is a measure of central tendancy of quantitative data and is widely used as a representative value for the data. The mean is the sum of all data and divided by the number of data, which implies the center of gravity of the data. The mean is expressed as μ (read mu), and the mean of (Data 4.1) is obtained as follows.

$$\text{Mean} = \mu = \frac{63 + 60 + 65 + 55 + 77}{5} = \frac{320}{5} = 64$$

- When n number of data is expressed as $x_1, x_2, \ldots, x_n$, mean is expressed by the following formula.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- In general, the mean is very appropriate as a representative value of the data, but when there is a very large or a small value in the data, it is greatly affected by this extreme value. In this case, a **median** can be used. The median is the value in the middle when the data are sorted in order. In (Data 4.1), the number of data is 5 which is an odd number, and the 3$^{\text{rd}}$ ($\frac{5+1}{2}$) number when data is sorted in ascending order is the median as follows.

(Data 4.1) is sorted in ascending order.

55 60 63 65 77

Median is the 3rd number in these sorted data whichi is 63.

- If the number of data is 6 which is an even number, how do we find the median? In this case, the median of the data is calculated as the average of the 3rd (= $\frac{6}{2}$) and 4th (= $\frac{6+2}{2}$) of the sorted data.

- Generally, a median is expressed as m, and, if the number of data is n, it is calculated as follows.

  1) Data are sorted in ascending order.
  2) Check whether the number of data n is an odd number or an even number.
  3) If n is odd, $m = (\frac{n+1}{2})^{th}$ data in sorted data

     If n is even, m = Average of $(\frac{n}{2})^{th}$ data $(\frac{n+2}{2})^{th}$ data in sorted data

- In order to see the overall distribution of the weight data, a stem and leaf plot or a histogram discussed in Chapter 3 can be considered, but a dot graph is more useful. In a dot graph, after obtaining the minimum and maximum values of data, the position of each data is calculated on the horizontal axis, and displayed as a dot.

- <Figure 4.1> is a dot graph for (Data 4.1). In proportion to the minimum value of 55 and the maximum value of 76, each data is displayed by a dot. The green line is the mean μ and the red line is the median m. In this data, the mean is located slightly to the right of the median because 77 of the data is located to the right of the other four data. That is, the mean is more sensitive to an extreme than the median.
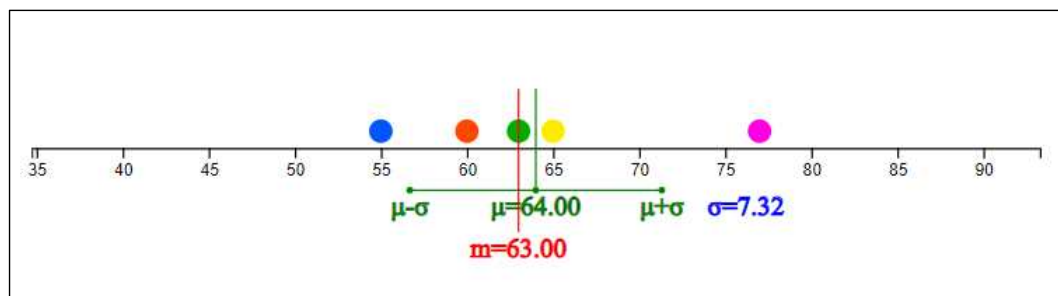
<Figure 4.1> Dot graph of weight data

- If there are lots of data, it is time-consuming and difficult to obtain the mean and median manually as above. Let's find the representative value of the data using 『eStatH』 software.

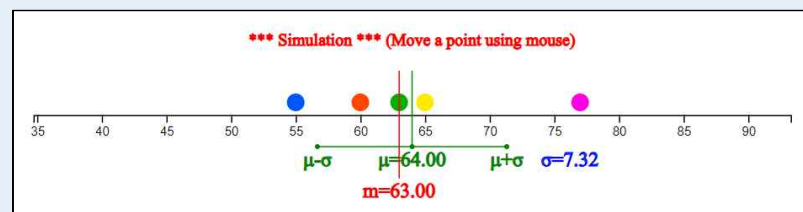| Practice 4.1	Using 『eStatH』, draw a dot graph for the weights of 5 students (Data 4.1) and find the mean and median.
Solution	• Using the QR on the left, select 'Dot Graph - Mean / Standard Deviation' from the 『eStatH』 menu, then a window like <Figure 4.2> appears. • Enter students' weight data in 'Data input'. (You can also copy and paste the data from the e-book) <Figure 4.2> Data input of weight data for a dot graph • When data are entered, the number of data, minimum, maximum, mean, and median are calculated immediately. If you click the [Execute] button, a dot graph appears as shown in <Figure 4.1> and the mean and median values are displayed. • Below <Figure 4.1>, a simulation window as shown in <Figure 4.3> appears. In this simulation window, you move a point with the mouse to see the changes in the mean and median. For example, if you drag the rightmost point and move it to the right, the mean changes but the median does not. That is, the median is not affected by the extreme points. <Figure 4.3> Simulation window to see a change in mean and median if you move a point

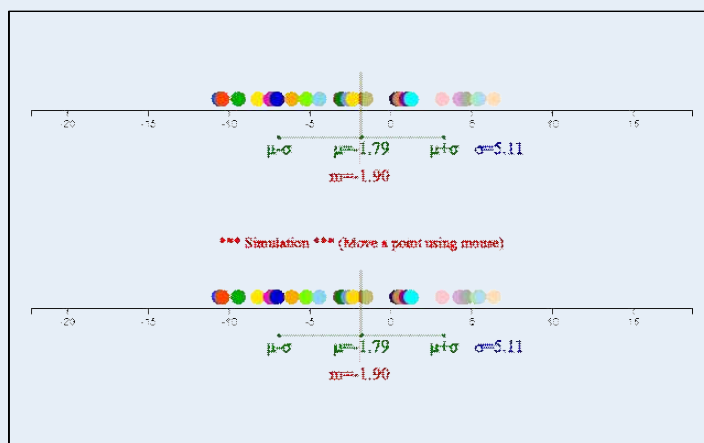| Practice 4.2	Using 『eStatH』, let's find the mean and median of the daily minimum temperature ([Practice 3.2]) in Seoul in February (Data 3.2).
	(Data 3.2) Daily minimum temperature in February 2021 in Seoul (unit eegree in Celcius) −2.3 −8.2 −9.4 −7.4 −4.4 4.3 −2.6 5.4 −6.1 −1.5 1.3 0.6 1.0 6.4 −5.2 −7.0 −10.4 −10.6 −7.1 5.5 4.7 0.4 −3.1 −3.0 0.7 0.5 4.3 3.2
Solution	• If you select 'Dot Graph – Mean/Standard Deviation' from the 『eStatH』 menu that appears using the QR on the left, the data input window as shown in <Figure 4.4> appears. <Figure 4.4> Temperature data input for a dot graph • When the daily minimum temperature data is entered in 'Data input' (you can copy and paste the data from the e-book), as shown in <Figure 4.4>, it show immediately that the number of data is 28, mean −1.79, median −1.90, minimum −10.6, maximum 6.4 degrees. • If you click the [Execute] button, a dot graph as shown in <Figure 4.5> appears and the mean (μ) and median (m) are displayed. Below this dot graph, a simulation window appears where you can change a point with the mouse and observe the changes in the mean and median values. <Figure 4.5> Dot graph of daily minimum temperature and its simulation window

Practice 4.2 Solution (Continued)	• Looking at this dot graph, it can be seen that there is almost no difference between the mean and the median due to the absence of extreme value.

Exercise 4.1	The following is data on the length of bicycle-only roads by 25 administrative districts in Seoul as of 2019 ([Exercise 3.1]). Use 「eStatH」 to draw a dot graph and to find and analyze the representative values of data. (Data 3.3) Length of bicycle-only roads by 25 administrative districts in Seoul as of 2019 (unit km) 24 15 23 20 30 24 7 8 7 12 28 27 19 35 41 42 11 8 37 13 20 29 53 93 42

Exercise 4.2	The following is data on the maximum wind speed of typhoons that passed through Korea in 2020 ([Exercise 3.2]). Use 「eStatH」 to draw a dot graph and to find and analyze the representative values of data. (Data 3.4) Maximum wind speed of typhoons that passed through Korea in 2020 (unit m/sec) 40 22 21 29 19 22 24 45 49 55 24 27 29 35 19 24 35 40 56 24 21 43 18

A. Calculation of mean using frequency table

☞ Think	Assume that a frequency table of the academic achievement test scores of a middle school class is given as follows. [Table 4.1] Frequency table of the academic achievement test scores of a middle school

Class	Frequency
$60 \leq$ ~ < 70	2
70 ~ 80	10
80 ~ 90	15
90 ~ 100	3
Total	30

Explore	How do we find the mean using this frequency table?

- When a frequency table is given rather than the raw data, the mean can be obtained approximately as follows using the middle values of each c;ass interval.

- First, find the middle value of each class interval. Then, it is assumed that each class has the middle value as many as the frequency, and the mean is obtained using this approximated data.

[Table 4.2] Approximated data using the middle value of each class interval in the academic achievement test scores

Class	Middle value	Frequency	Approximated data
60 $\leq$ ~ $<$ 70	65	2	65 65
70 ~ 80	75	5	75 75 75 75 75
80 ~ 90	85	10	85 85 85 85 85 85 85 85 85 85
90 ~ 100	95	3	95 95 95
Total		20	

- Mean is calculated as follows.

$$\text{Mean} = \frac{65+65+75+75+75+75+75+85+85+85+85+85+85+85+85+85+85+95+95+95}{20}$$

$$= \frac{65\times2 + 75\times5 + 85\times10 + 95\times3}{20}$$

$$= \frac{1640}{20} = 82$$

- Using 'Frequency Distribution Polygon – Relative Frequency Comparison' of 『eStatH』, the approximate mean of the frequency table can be obtained as shown in <Figure 4.6>. After entering the left value of the class interval and 'Frequency 1', click the [Execute] button.

<Figure 4.6> Mean calculation using a frequency table

4.2   Measure of Dispersion – Standard Deviation

☞ Think	The quiz scores (out of 10) of five middle school students are as follows.
	(Data 4.2) The quiz scores (out of 10) of five middle school students
	6 8 7 4 10
Explore	Is there a way to measure how scattered these data are?

- The degree to which data are scattered is called a dispersion. A simple measure of the dispersion is a **range** which is the maximum minus the minimum.

  Range = max - min

  In (Data 4.2), the maximum value is 10 and the minimum value is 4, so the range is 22.

  Range = 10 - 4 = 6

- Since the range is too sensitive to extreme values, a variance or a standard deviation is generally used to measure the dispersion. The **variance** is obtained by squaring the distance between each data value and the mean, and dividing it by the number of data. Therefore, when the data are scattered far from the mean, the variance is large, and when the data are clustered around the mean, the variance is small. The variance is expressed as σ^2 (read as sigma squared).

- The mean of the data in (Data 4.2) is as follows.다.

  $$\text{Mean} = \mu = \frac{6+8+7+4+10}{5} = \frac{35}{5} = 7$$

- The variance is calculated by squaring the distances from the mean to each data value to find the sum, and then finding the mean. That is, it is the average of squared distances.

  $$\text{Variance} = \sigma^2 = \frac{(6-7)^2 + (8-7)^2 + (7-7)^2 + (4-7)^2 + (10-7)^2}{5} = \frac{20}{5} = 4$$

- When n number of data is expressed as $x_1, x_2, \ldots, x_n$ and the mean is expressed as μ, the variance can be expressed by the following formula.

$$\text{Variance} \qquad \sigma^2 \; = \; \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \qquad (n: \text{ number of data})$$

- The **standard deviation** is defined as the square root of the variance and denoted by σ. The variance is not easy to interpret practically because it is the average of the squared distances, but the standard deviation is the square root of the variance, so it can be interpreted as a measure of the average distance between each data value and the mean.

$$\text{Standard deviation} \qquad \sigma \; = \; \sqrt{\sigma^2}$$

The standard deviation of (Data 4.2) is $\sigma \; = \; \sqrt{\sigma^2} \; = \; \sqrt{4} \; = 2$.

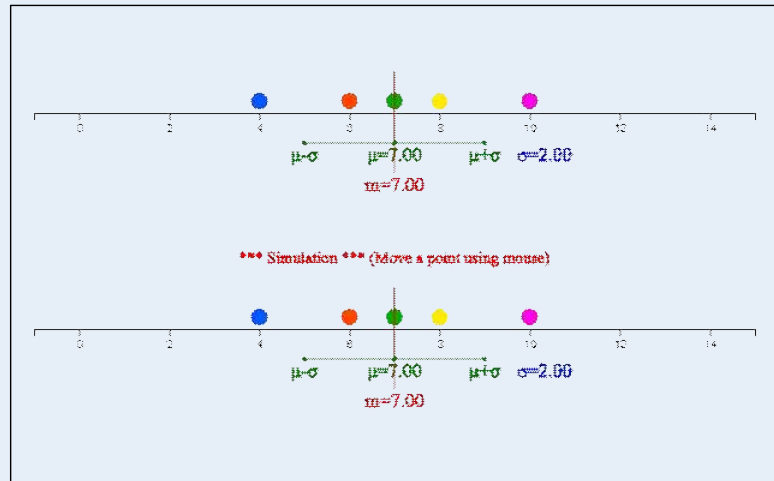| Practice 4.3	Using 「eStatH」, draw a dot graph for the quiz scores of 5 sample students (Data 4.2) and find the mean and standard deviation.
Solution	• Using the QR on the left, select 'Dot Graph – Mean / Standard Deviation' from the 「eStatH」 menu. Then a window like <Figure 4.7> appears. • Enter students' quiz scores in 'Data input'. (You can also copy and paste the material from the e-book) **Dot Graph - Mean/StdDev** `Menu` **[Enter Data]** `6 8 7 4 10` Sorted Data `4,6,7,8,10` Number of Data n 5 Minimum min 4.00 Total 35.00 Maximum max 10.00 Mean μ 7.00 Range $range$ 6.00 Median m 7.00 Variance σ^2 4.00 Mode $mode$ Std Deviation σ 2.00 `Execute` <Figure 4.7> Quiz data input for a dot graph • When the data are entered, the number of data, minimum, maximum, mean, median, variance and standard deviation are calculated. If you click the [Execute] button, a dot graph as shown in <Figure 4.8> appears and the mean, median, standard deviation, and a line of mean ± standard deviation are displayed. • Using the simulation window below the figure, you can check the change in the standard deviation by moving a data point with the mouse. The standard deviation is also affected by an extreme point.

Practice 4.3 Solution (continued)	 <Figure 4.8> Dot graph with a line of mean ± standard deviation

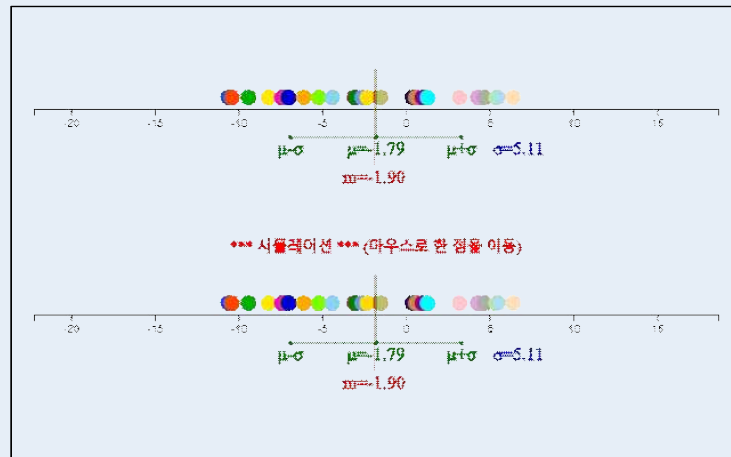| Practice 4.4	Using 『eStatH』, let's draw a dot graph for the daily minimum temperature ([Practice 3.2]) in Seoul in February (Data 3.2) and find the mean and standard deviation.
	
	(Data 3.2) Daily minimum temperature ([Practice 3.2]) in Seoul in February in 2021 (degree in Celsius)
	
	−2.3 −8.2 −9.4 −7.4 −4.4 4.3 −2.6 5.4 −6.1 −1.5
	  1.3 0.6 1.0 6.4 −5.2 −7.0 −10.4 −10.6 −7.1 5.5
	  4.7 0.4 −3.1 −3.0 0.7 0.5 4.3 3.2

Solution	• If you select 'Dot Graph − Mean / Standard Deviation' from the 『eStatH』 menu that appears using the QR on the left, the data input window as shown in <Figure 4.9> appears. <Figure 4.9> Temperature data input for a dot graph • When data are entered, the number of data, minimum, maximum, mean, median, variance and standard deviation are calculated. If you click the [Execute] button, a dot graph as shown in <Figure 4.10> appears and the mean, median, standard deviation, and a line of mean ± standard deviation are displayed.

Practice 4.4 Solution (continued)	• Using the simulation window below the figure, you can check the change in the standard deviation by moving a point with the mouse. The standard deviation is also affected by an extreme point.
	 <Figure 4.10> Dot graph of daily minimum temperature and a simulation window

Exercise 4.3	The following is data on the length of bicycle-only roads by 25 administrative districts in Seoul as of 2019 ([Exercise 3.1]). Use 「eStatH」 to draw a dot graph and to find and analyze the mean and standard deviation of the data.
	(Data 3.3) Length of bicycle-only roads by 25 administrative districts in Seoul as of 2019. (unit km) 24 15 23 20 30 24 7 8 7 12 28 27 19 35 41 42 11 8 37 13 20 29 53 93 42

Exercise 4.4	The following is data on the maximum wind speed of typhoons that passed through Korea in 2020 ([Exercise 3.2]). Use 「eStatH」 to draw a dot graph and find and analyze the mean and standard deviation of data.
	(Data 3.4) Maximum wind speed of typhoons that passed through Korea in 2020 (unit m/sec) 40 22 21 29 19 22 24 45 49 55 24 27 29 35 19 24 35 40 56 24 21 43 18

A. Calculation of standard deviation using frequency table

☞ Think	Assume that the frequency table of the academic achievement test scores of a middle school class is given as follows.

[Table 4.3] Frequency table of the academic achievement test scores of a middle school

Class	Frequency
$60 \leq \sim < 70$	2
70 ~ 80	10
80 ~ 90	15
90 ~ 100	3
Total	30

Explore	How to find the standard deviation of the data in this frequency table?

- In the previous section, when a frequency table was given rather than the raw data, the mean was approximated using the middle value of each class interval. The standard deviation is calculated in a similar way.
- First, find the middle value of each class. Then, it is assumed that each class has the middle value as many as the frequency, and the average is obtained using this approximate data.

[Table 4.4] Approximated data using the middle value of each class interval in the academic achievement test scores

Class	Middle value	Frequency	Approximated data
$60 \leq \sim < 70$	65	2	65 65
70 ~ 80	75	5	75 75 75 75 75
80 ~ 90	85	10	85 85 85 85 85 85 85 85 85 85
90 ~ 100	95	3	95 95 95
Total		20	

- Mean is calculated as follows.

$$\text{Mean} = \frac{65+65+75+75+75+75+75+85+85+85+85+85+85+85+85+85+95+95+95}{20}$$

$$= \frac{65\times2 + 75\times5 + 85\times10 + 95\times3}{20}$$

$$= \frac{1640}{20} = 82$$

- The variance and standard deviation are calculated in a similar way.

$$\text{Variance} = \frac{(65-82)^2\times2 + (75-82)^2\times5 + (85-82)^2\times10 + (95-82)^2\times3}{20}$$

$$= \frac{1420}{20} = 71$$

Standard deviation = $\sqrt{71}$ = 8.43

- Using 'Frequency Distribution Polygon – Relative Frequency Comparison' of 『eStatH』, the approximate mean and standard deviation of the frequency table can be obtained as shown in <Figure 4.6>. After entering the left value of the class interval and 'Frequency 1', click the [Execute] button.

<Figure 4.11> Standard deviation calculation using a frequency table

4.3 Covariance and Correlation Coefficient

☞ Think	The height and weight of 7 male middle school students were investigated as follows.

(Data 4.3) Height and weight of 7 male middle school students

	1	2	3	4	5	6	7
Height (cm)	162	164	170	158	175	168	172
Weight (kg)	54	60	64	52	65	60	67

Explore	Is there a measure to determine the correlation between two quantitative variables?

- Just as the variance is used as a measure of dispersion in one quantitative variable, the following covariance is used in two quantitative variables. When n number of x and y data are expressed as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, and the mean is expressed as (μ_x, μ_y), the **covariance** σ_{xy} can be expressed by the following formula.

$$\text{Covariance} \qquad \sigma_{xy} \ = \ \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y) \qquad (n\text{: number of data})$$

- Covariance implies the total average of the values obtained by multiplying the x-axis distance and the y-axis distance between each data point and the mean point (μ_x, μ_y) of data. Therefore, if there are many points on the upper right and lower left of the mean point, the covariance has a positive value, indicating a positive correlation. If there are many points on the upper left and lower right of the mean point, the covariance has a negative value, indicating a negative correlation. However, since covariance can increase in value depending on the unit of data, the following correlation coefficient is used as a measure of correlation.

$$\text{Correlation coefficient} \quad \rho \ = \ \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- The correlation coefficient is a variation of the covariance and can only have values between -1 and $+1$. When the correlation coefficient is close to $+1$, the two variables are said to have a strong positive correlation, and when the correlation coefficient is close to -1, it is said to have a strong negative correlation. When the correlation coefficient is close to 0, there is no correlation between the two variables.

Practice 4.5	Using 「eStatH」, calculate the covariance and correlation coefficient of the height and weight of 7 students.
Solution	- If you select 'Scatter Plot - Correlation Coefficient' from the 『eStatH』 menu using the QR on the left, a window like \<Figure 4.12\> appears. - Enter students' height in 'Enter X data' and their weight in 'Enter Y data'. (You can also copy and paste the material from the e-book)

Practice 4.5 Solution (continued)	

Scatterplot Menu

Main Title Hieght and weight of 7 male middle school students

y title Weight   **x title** Height

X Enter Data 162 164 170 158 175 168 172

Y Enter Data 54 60 64 52 65 60 67

Number of Data	n_x	7	n_y	7			
Mean	μ_x	167.00	μ_y	60.29			
Variance	σ_x^2	30.57	σ_y^2	27.06	Covariance	σ_{xy}	27.00
Std Deviation	σ_x	5.53	σ_y	5.20	Correlation Coefficient	ρ	0.94

Execute

<Figure 4.12> Calculation of covarince and correlation coefficient using height and weight data

- After the data input, click the [Execute] button, Then the number of data, mean, variance, standard deviation, covariance and correlation coefficient are calculated as <Figue 4.12> and a scatter plot is displayed.
- If you check the 'regression line' under the scatterplot, a regression line that explains the relationship between height and weight is drawn.
- As shown in <Figure 4.12>, the covariance of height and weight in (Data 4.3) is 27 and the correlation coefficient is 0.94, indicating a strong positive correlation.
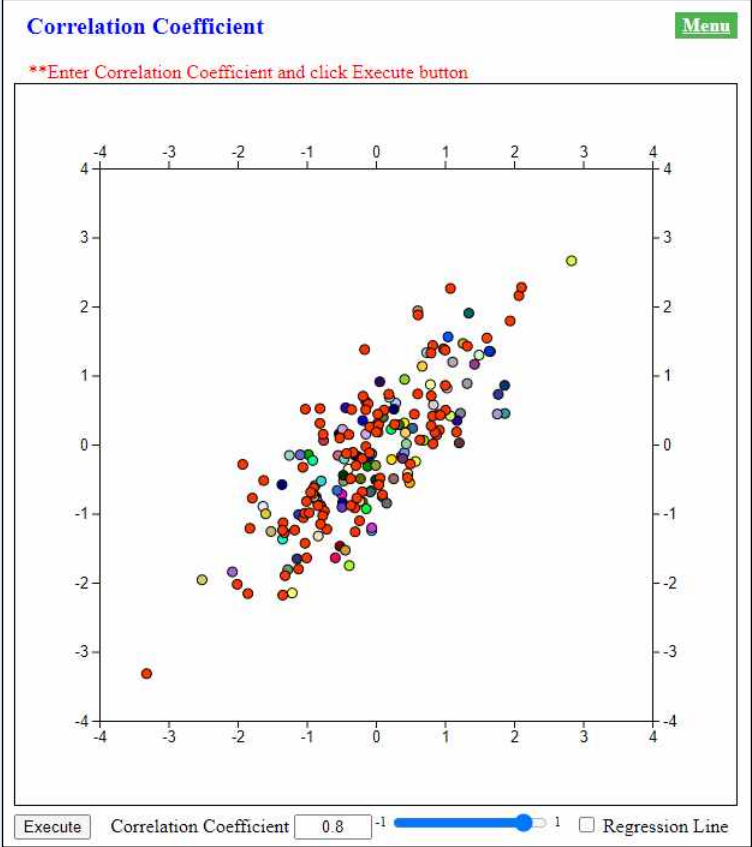
Exercise 4.5	The following are data on the weekly study hours and test scores of 10 middle school students. Find the covariance and correlation coefficient using 『eStatH』 and analyze what kind of correlation is.

(Data 4.4) Weekly study hours and test score of 10 students

	1	2	3	4	5	6	7	8	9	10
Study hours	10	25	15	16	20	5	18	21	12	20
Test score	75	95	82	85	97	65	87	88	76	90

- By using 『eStatH』, you can examine examples of data for various correlation coefficients.

Practice 4.6	Let's simulate various correlation coefficients using 『eStat H』.
Solution	• If you select 'Correlation Coefficient' from the QR on the left or 『eStatH』 menu, an initial scatter plot as shown in <Figure 4.13> appears. <Figure 4.13> Simulation window for a correlation coefficient • Change the 'Correlation Coefficient' under the initial scatter plot to the desired value and click the [Execute] button to display a scatter plot for the corresponding correlation coefficient. • If 'Regression Line' is checked, a regression line representing the points appears.

• If the correlation is strong, a straight line that can explain the relationship between the variables is obtained, which is called a regression line. A detailed explanation of the regression line is covered in university level of statistics.

Practice 4.7	Using 『eStatH』, make a point on the plane and observe the correlation coefficient and regression line while moving.
Solution	• If you select 'Correlation Coefficient – Regression Line Experiment' from the QR on the left or the 『eStatH』 menu, a screen for testing the correlation coefficient and regression line as shown in <Figure 4.14> appears.

<Figure 4.14> Simulation window of a
correlation coefficient and a regression line

• If you place a dot on this blank screen with the mouse, the regression line and correlation coefficient appear as shown in <Figure 4.15>. You can observe the change of the regression line and the correlation coefficient by moving the point after clicking it with the mouse.

<Figure 4.15> Example of dots and its
regression line and correlation coefficient