1. Statistics

- Examine what statistics are.
- Examine the classification of data.
- Examine statistical software.

1.1 What is Statistics?

☞ Think	The height of 10 students in our class was measured to see how tall they were. The measured height data were recorded on paper as follows. [Table 1.1] The height of 10 students in our class (unit cm) 165 163 170 175 155 167 157 153 161 162
Explore	- How can I find out where my height (165cm) is located among students in my class in descending order of height?

- Students' heights are called numerical **data**. In this example, the data was generated as many as the number of students. In our society, a large amount of data were generated and the number of data is expected to increase further with the development of computer technologies in the future.
- In order to understand the relative position of my height among students, arrange the data in descending order from taller to smaller as shown in [Table 1.2].

[Table 1.2] Height data in descending order (unit cm)

175	170	167	165	163	162	161	157	155	153

From this data, it is easy to tell that my height of 165cm is the 4th from the top.

- **Statistics** is the process of collecting, organizing, and summarizing data in this way and obtaining useful information by analysis. A country needs statistics such as the number of people, the number of households, and the amount of food in order to solve eating, clothing and sheltering of its people. Such statistics have been recorded since the time of Egyptian civilization 5,000 years ago.
- As the modern information society progresses, the size of data has grown tremendously, which is called Big Data. Typical examples of big data are as follows.

  - Google's search history data
  - SNS data from your smartphone
  - Records of access to and out of the Internet

- Telecommunication company's call log data

- In the information society of the future, the success or failure of an individual, company, or country can be different depending on how they make scientific decisions with reasonable statistical informations on big data. There are many examples of statistics applied to reality by analyzing big data.

  - Predict the number of car models sold in the United States next month by surveying Google's search engine questions about car purchases.
  - An online shopping mall analyzed customers' web records to find out what products the member customers have and what products they were interested in, and then customized advertisements tailored to each customer, resulting in increased sales.
  - An oil exploration company increased the success rate of oil drilling by analyzing terabytes of geological data.
  - A university in the United States analyzed the information that students click on the system in an online class to monitor each student's learning performance, suggested class content for each level tailored to the student's level of understanding, and suggested subjects to take in the future for each student. As a result, the degree acquisition rate by major improved significantly.

1.2 Data Classification

☞ Think	If the gender of the 10 students in our class is investigated, they are 'male' or 'female', and if the height is measured, it is 160cm, etc. Gender and height data have different attributes.
Explore	1) Gender and height have different attributes. Is there an appropriate way to classify data? 2) Does the analysis method differ depending on the type of data?

- Data are values that observed or measured the attributes of objects of interest. The attribute of such an object or event is called a **variable**. For example, if the gender and height of an elementary school were measured, there were two variables (gender and height). Gender measurements are in the form as male, female, female, male, ... and height measurements are

160cm, 155cm, 158cm, 165cm, ...

- Variable such as 'gender' is classified into **qualitative data**, and variable such as height is classified into **quantitative data**. The reason for classifying data is that processing and analyzing methods are different depending on the type of data. Chapter 2 discusses visualization of qualitative data, and Chapter 3 discusses visualization of quantitative data.
- Qualitative variable such as gender usually have several categories, which are also called a **categorical variable.** For a categorical variable, the frequency of each category is investigated, a frequency table is created, and a bar graph, pie graph, and band graph are drawn for analyis (Chapter 2). Table 1.3 is an example of a frequency table for gender in a class. It shows that there are 6 'male' and 4 'female' students.

[Table 1.3] Frequency table for gender in a class

Gender	Number of students
Male	6
Female	4

- Qualitative data such as words are analyzed by examining the frequency of each word to form a word cloud.
- For quantitative data, the mean and standard deviation are calculated and visualized using stem and leaf plot, histogram, and dot graph (Chapter 3).

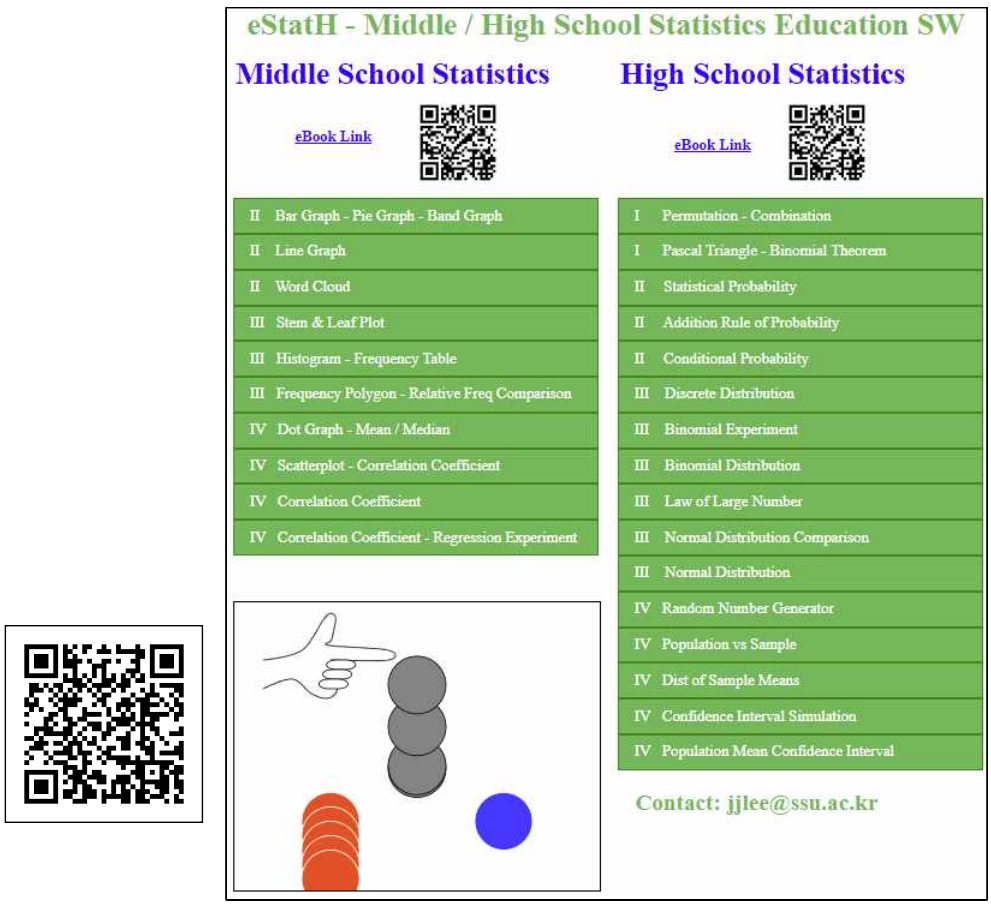
1.3 『eStat』 Software

☞ Think	[Table 1.2] arranged the height data in descending order on a paper
Explore	Is there any way to use a computer? If yes, which software should I use?

- Small class data can be manually organized and analyzed, but if there are more data, it is impossible to analyze data without the help of a computer. In particular, **statistical software** with many professional analysis programs is essential for big data analysis. Statistical software such as SAS, SPSS, and R are now widely used for data analysis. However, these software are not easy for beginners to learn, and SAS and SPSS are commercial and are very expensive. In addition, these software are focused

on data analysis and there are few module required for statistical education.

- 『eStat』 is a free educational software designed to teach statistics easily from elementary school students to college students and general public. Given the data, you can draw a graph with just a click of the mouse, and experience a dynamic graph. 『eStat』 includes many examples from textbooks of each level, and because it is web-based, users can use it anytime, anywhere with a PC, tablet, or smartphone. 『eStat』 is provided free of charge and supports multiple languages, and currently 20 languages including Korean and English are available.
- 『eStat』 system works 100% well with Google's Chrome among many web browsers currently in use. In browsers such as MS Edge, 『eStat』 works, but some functions may not work.
- This book is focused on modules in 『eStatH』 developed for elementary and middle school students in 『eStat』. Open the web browser Chrome, enter www.estat.me in the address bar and click the icon 『eStatH』, then the screen as shown in <Figure 1.1> appears.



<Figure 1.1> 『eStatH』 menu

Exercise

1.1 Which of the following statements about statistics is correct? (Answer ②)
① Studying changes in the weather
② Collect and organize data for analysis
③ Study chemical reactions
④ Electronic production research

1.2 Which of the following is qualitative data? (Answer ③)
① Height ② Weight ③ Education ④ Age

1.3 Which of the following is quantitative data? (Answer ④)
① Gender ② Education ③ Marital status ④ Age

1.4 Which of the following is not a statistical package? (Answer ④
① SPSS ② SAS ③ eStat ④ MS Word

2. Visualization of Qualitative Data

In the case of categorical data, the following graphs are drawn and analyzed.

- bar graph
- pie graph
- rainbow graph
- band graph
- line graph

The data in the form of words are analyzed by drawing a word cloud.

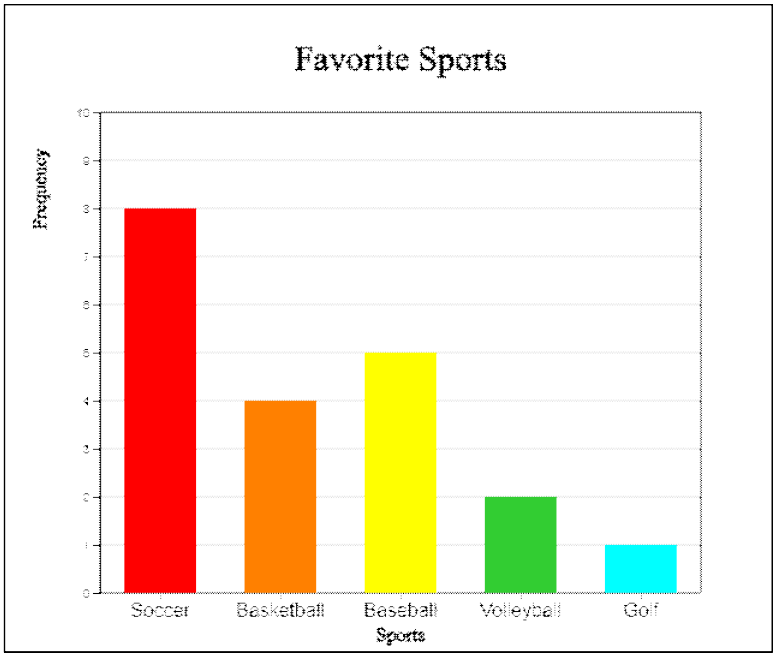
2.1 Bar / Pie / Rainbow / Band Graph

☞ Think	The frequency table that surveyed 20 students in an elementary school class for their favorite sports is as follows.
Explore	What are some graphs that can be easily observed what a favorite sport is like?

[Table 2.1] Frequency table for favorite sports

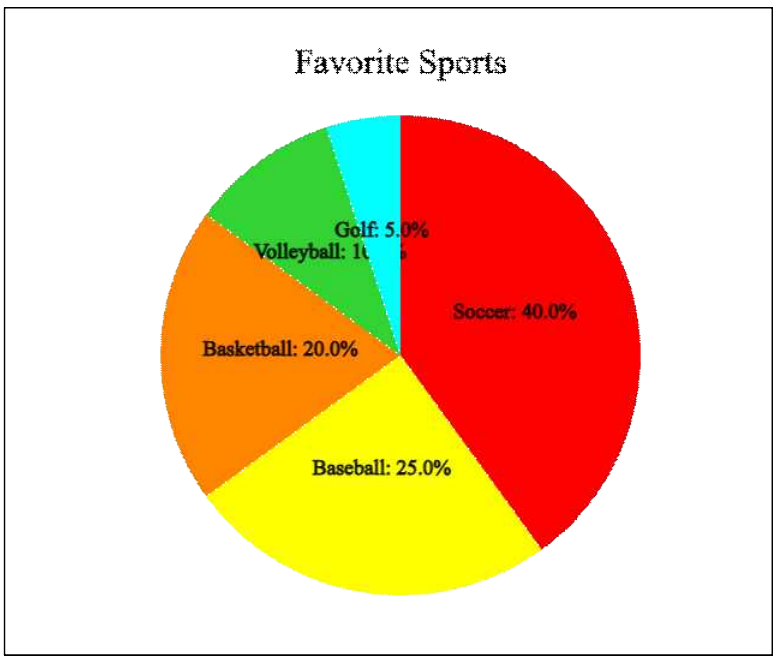
Sports	Number of students
Soccer	8
Basketball	4
Baseball	5
Volleyball	2
Golf	1

- When you ask your students about their favorite sports, they say, 'soccer', 'basketball', 'baseball', ... You can get answers like this, which is called **qualitative data**. Most qualitative data are referred to as categorical data because their values represent categories. Qualitative data can be summarized as a frequency table by counting the number of students in each category, and examine their characteristics using a bar graph, a pie graph, a rainbow graph and a band graph.
- A **bar graph** shows the frequency of each category of data at the height of a rectangular bar. Spacing between the bars is used to emphasize that it is categorical data (<Figure 2.1>).



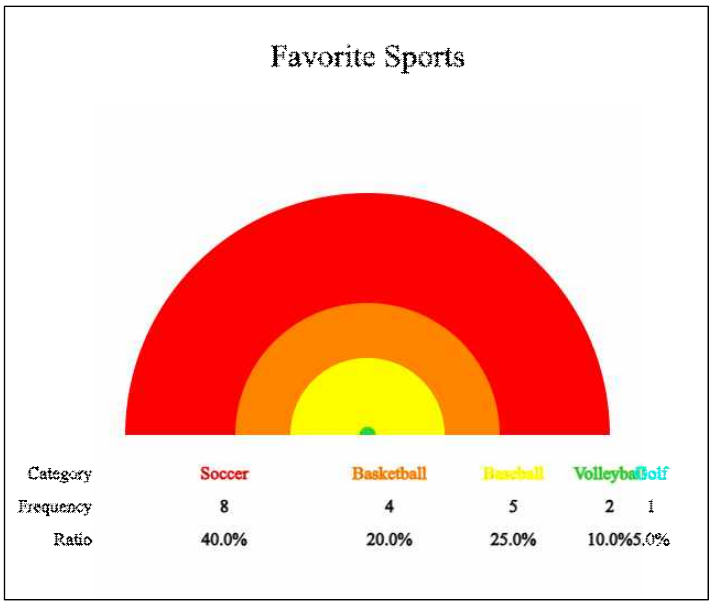
<Figure 2.1> Bar graph for favorite sports

- A **pie graph** is a graph in which the frequency of each category in the wholel data is divided into pie slices. For better comparison of proportions, draw the circle pieces in the order of the largest category in a clockwise direction.



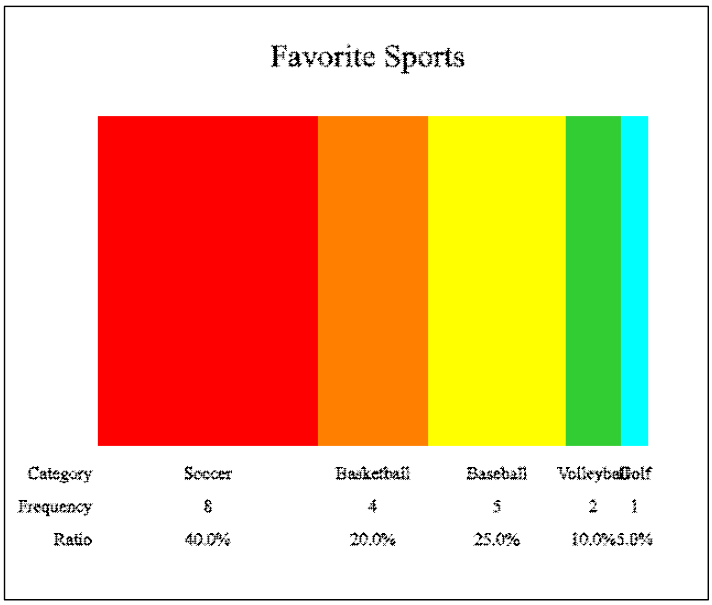
<Figure 2.2> Pie graph for favorite sports

- A rainbow **graph** is a modified form of a pie graph and is a graph in which the frequency of each category is divided into raibow pieces and displayed in the total data.



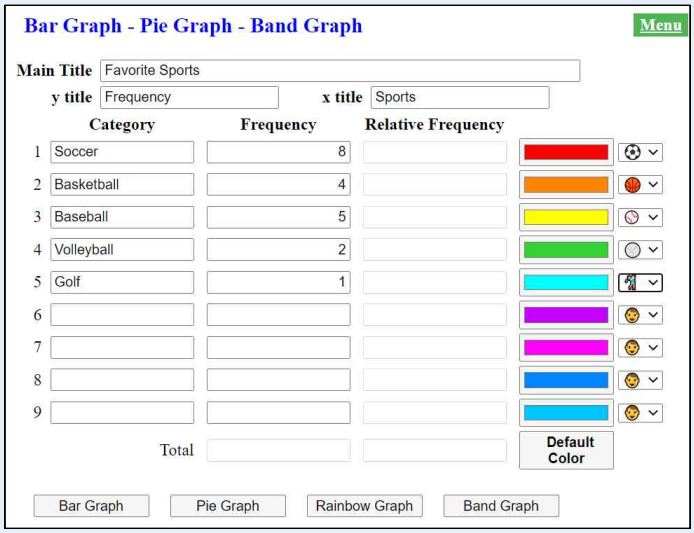
<Figure 2.3> Rainbow graph for favorite sports

- A **band graph** is a modified form of a pie graph, and is a graph in which the frequency of each category is divided into rectangle pieces and displayed in the total data. It is also possible to sort in the order of the largest value of the category, but to distinguish it from the pie graph, the rectangle pieces are drawn in the order of the categories of the data in 『eStatH』.

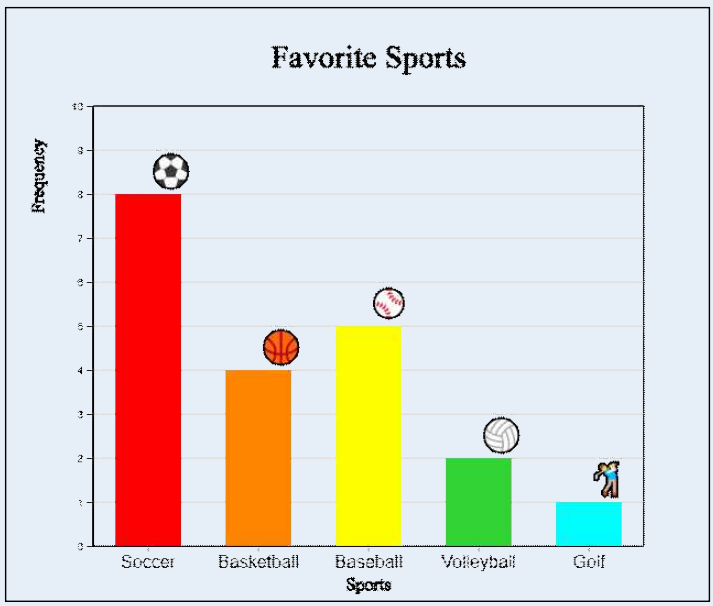


<Figure 2.4> Band graph for favorite sports

Practice 2.1	Using 『eStatH』, draw a bar graph, pie graph, and band graph for favorite sports and observe which sports students like the most.
Solution	• If you select 'Bar Graph - Pie - Band Graph' from the 『eStatH』 menu using the QR on the left, the data input window as shown in <Figure 2.5> appears. • Enter the desired title for 'Main Title', 'y title', and 'x title' and enter 'Category' and 'Frequency'. • Click the [Bar Graph] button to display a bar graph as in <Figure 2.1>. You can also draw the bar graph again by selecting the color of the desired category. If you select the icon for each category on the far right and check 'Emoji' under the graph, a bar graph like <Figure 2.6> appears.



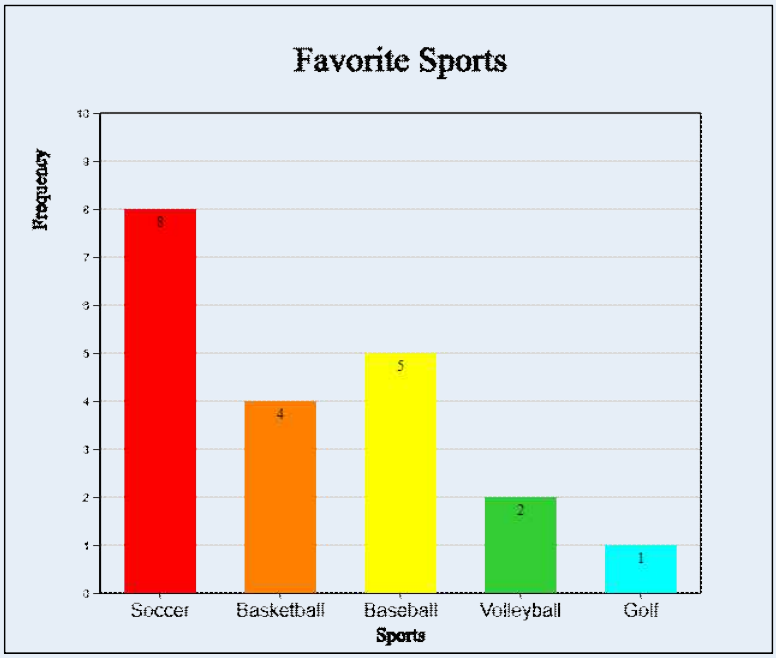
<Figure 2.5> Data input for favorite sports



<Figure 2.6> Bar graph with Imoji icons for each category

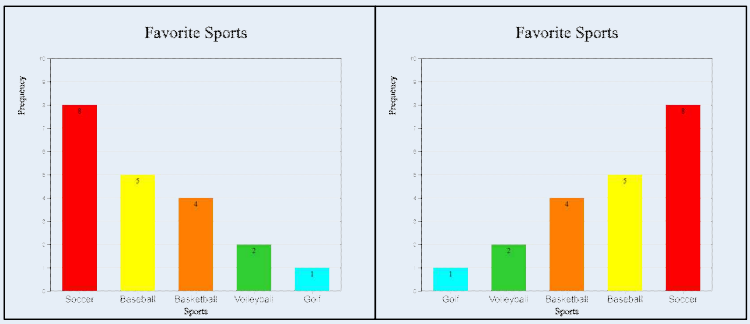
Practice 2.1 Solution (Continued)

- In a similar way, the frequency can be displayed on the bar as shown in <Figure 2.7>.



<Figure 2.7> Bar graph with frequencies

- By using the (Soring) option under the graph, you can sort the bars in descending or ascending order of frequency as in <Figure 2.8> or <Figure 2.9>.



<Figure 2.8> Bar graph in descending order of frequencies

<Figure 2.9> Bar graph in ascending order of frequencies

- If you click 'Pie Graph', 'Rainbow Graph', 'Band Graph', you can see the graphs such as in <Figure 2.2>, <Figure 2.3> and <Figure 2.4> respectively.

Practice 2.2

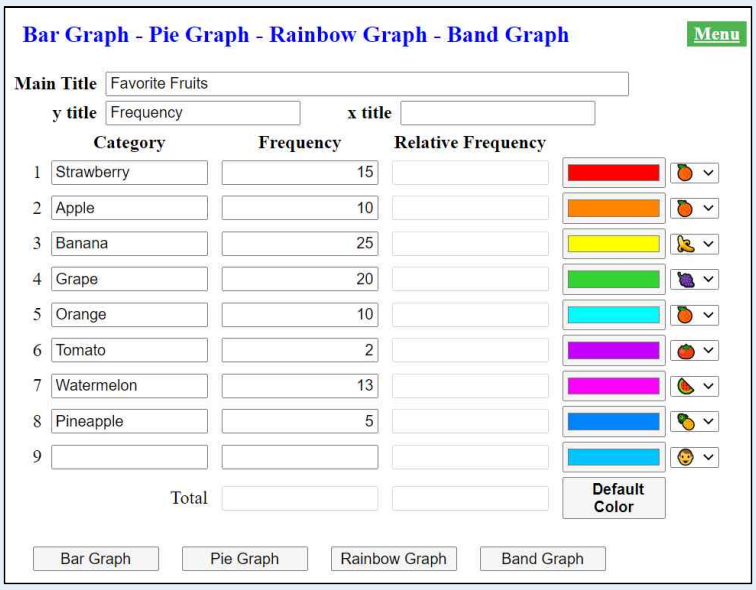
The fruits that elementary school students liked were as follows. Let's draw a bar graph using 『eStatH』.

[Table 2.2] Fruits that students liked

Fruits	Number of students
Strawberry	15
Apple	10
Banana	25
Grape	20
Orange	10
Tomato	2
Watermelon	13
Pineapple	5

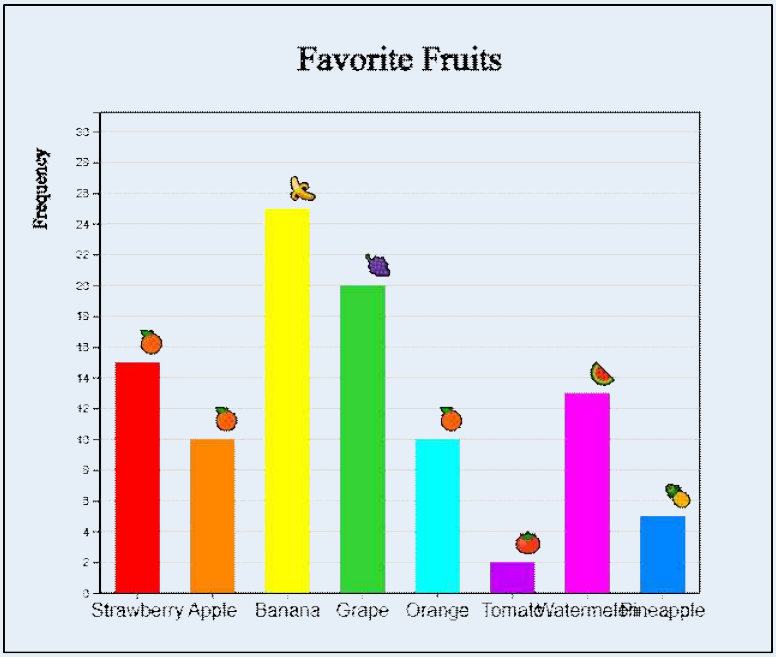
Practice 2.2 Solution

- If you select 'Bar Graph – Pie Graph – Band Graph' from the 『eStatH』 menu using the QR on the left, the data input window as shown in <Figure 2.10> appears.
- Enter the desired title for 'Main Title', 'y title', and 'x title' and enter 'Category' and 'Frequency'. Select each fruit imoji icon on the far right.
- Click the [Bar Graph] button and when the graph appears, check the 'Emoji' below the graph. You can also draw the bar graph again by selecting the color of the desired category.



<Figure 2.10> Data input for favorite fruits

Practice 2.2 Solution



<Figure 2.11> Bar graph for favorite fruits with imoji

Exercise 2.1

According to the United Nations 2018 estimates, World largest 10 cities (mixture of city, metropolitan and urban area) are as follows. Draw a bar graph, pie graph, rainbow graph and band graph using 『eStatH』 to find out the characteristics.

[Table 2.3] World largest 10 cities (mixture of city, metropolitan and urban area)

City	Population (unit million)
Tokyo (Japan)	37.4
Delhi (India)	28.5
Seoul (Korea)	25.7
Shanghai (China)	25.6
San Paulo (Brazil)	21.7
Mexico City (Mexico)	21.6
Cairo (Egypt)	20.1
Mumbai (India)	20.0
Beijing (China)	19.6
Dhaka (Bangladesh)	19.6

Exercise 2.2	A summary of the survey of prospective jobs of elementary school students is as follows. Use 『eStatH』 to draw a bar graph, a pie graph, rainbow graph and band graph to identify the characteristics.

[Table 2.4] Prospective jobs of elementary school students

Job	Nomber of Students
Movie star	45
Doctor	30
Teacher / Professor	60
Lawer	24
Pro Gramer	10
Sportsman	43
Police	26

2.2 Line Graph

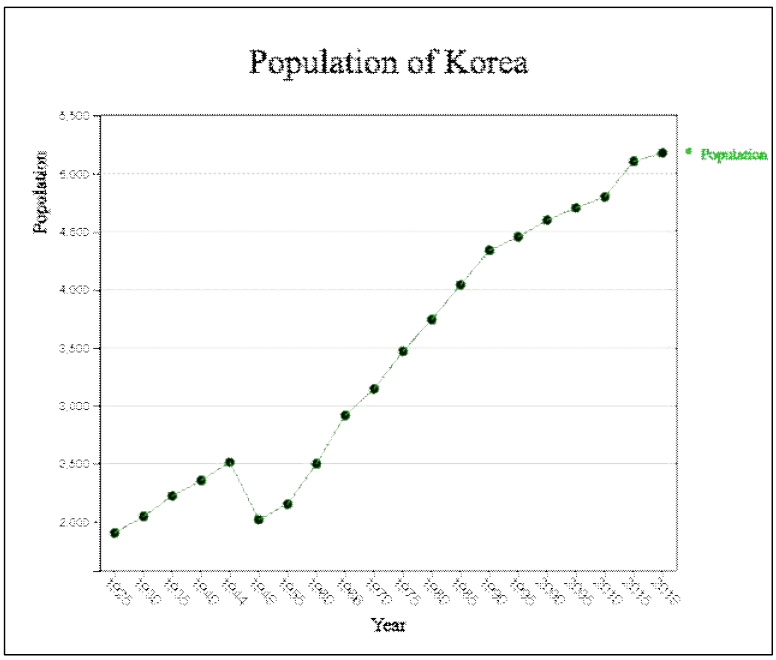
☞ Think	The following table shows the population of Korea, surveyed at approximately five-year intervals from 1925 to 2019.

[Table 2.5] Population of Korea

Year	Population (unit 10000)
1925	1902
1930	2044
1935	2221
1940	2355
1944	2512
1949	2017
1955	2150
1960	2499
1966	2916
1970	3144
1975	3468
1980	3741
1985	4042
1990	4339
1995	4455
2000	4599
2005	4704
2010	4799
2015	5106
2019	5178

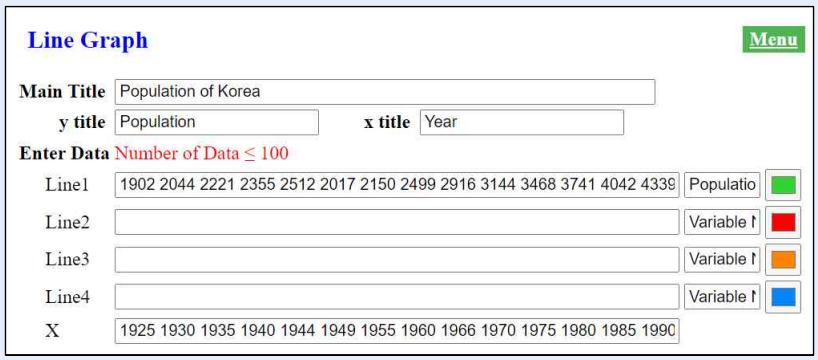
Explore	What kind of graph can easily observe the data measured over time?

- Data measured over time is often observed using a **line graph.** In a line graph, each data is displayed as a dot on the XY plane, with the time such as month or year as the X axis and the value of the other variable as the Y value, and then connected with a line. It is similar to a bar graph, but the change in data over time can be observed.
- Observing the line graph of Korea's population, it can be seen that it was increased from 1925 to 1944, and then decreased significantly in 1949 and 1954 due to the Pacific War and the Korean War. After that, the population continued to increase significantly during the baby boom period in the late 1950s, but the trend of increase became moderate after 1990, and this trend has become more gentle after 2015.



<Figure 2.12> Line graph of Korea's population

Practice 2.3	Using 『eStatH』, draw and observe a line graph for the Korean population in [Table 2.5].
Solution	• If you select 'Line Graph' from the 『eStatH』 menu using the QR on the left, a data input window as shown in <Figure 2.13> appears. • Enter the 'main title', 'y title', and 'x title' as shown in the figure, enter the number of population by year in 'Line 1', the variable name next to it, and year data in 'X'. • If you click the [Execute] button, a line graph as shown in <Figure 2.12> appears. You can also draw a line graph again by selecting the color of each line variable.



<Figure 2.13> Data input for Line Graph

- When there are multiple variables observed at the same time, the line graph can draw multiple lines simultaneously on one graph.

Practice 2.4

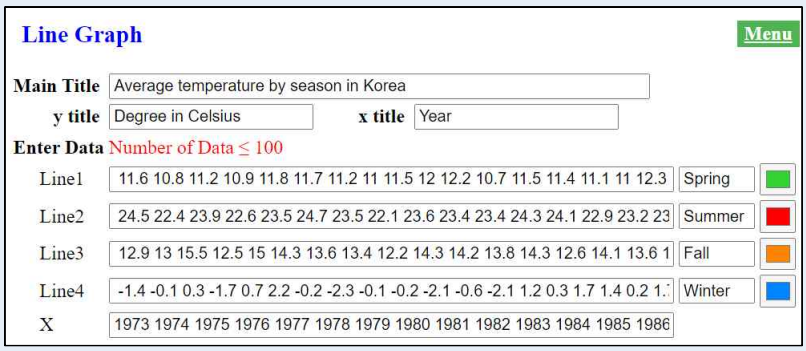
Using 『eStatH』, draw and observe the line graph for the average temperature of spring, summer, fall, and winter season in Korea such as in [Table 2.6] on one graph.

[Table 2.6] Average temperature of each season in Korea
(unit: degree in Celsius)

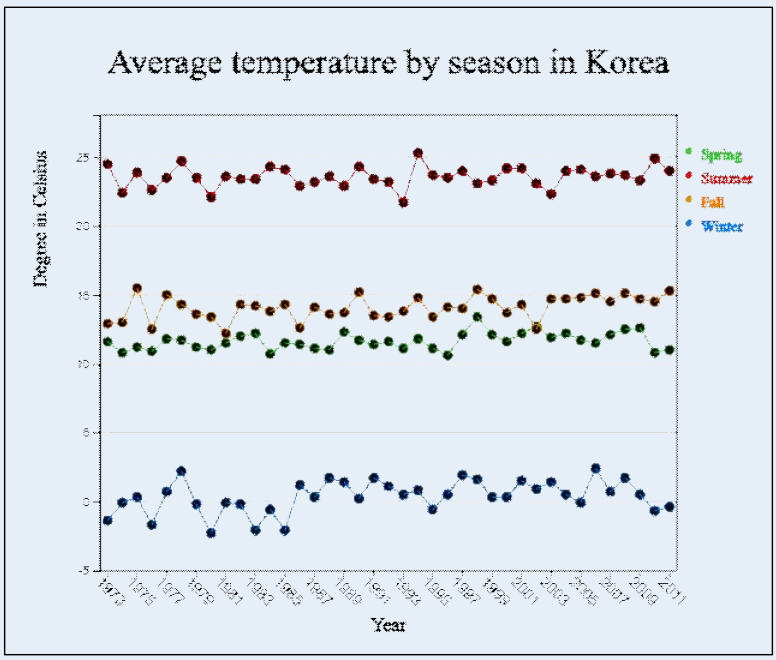
Year	Spring	Summer	Fall	Winter
1973	11.6	24.5	12.9	-1.4
1974	10.8	22.4	13	-0.1
1975	11.2	23.9	15.5	0.3
1976	10.9	22.6	12.5	-1.7
1977	11.8	23.5	15	0.7
1978	11.7	24.7	14.3	2.2
1979	11.2	23.5	13.6	-0.2
1980	11	22.1	13.4	-2.3
1981	11.5	23.6	12.2	-0.1
1982	12	23.4	14.3	-0.2
1983	12.2	23.4	14.2	-2.1
1984	10.7	24.3	13.8	-0.6
1985	11.5	24.1	14.3	-2.1
1986	11.4	22.9	12.6	1.2
1987	11.1	23.2	14.1	0.3
1988	11	23.6	13.6	1.7
1989	12.3	22.9	13.7	1.4
1990	11.7	24.3	15.2	0.2
1991	11.4	23.4	13.5	1.7
1992	11.6	23.2	13.4	1.1
1993	11.1	21.7	13.8	0.5
1994	11.8	25.3	14.8	0.8
1995	11.1	23.7	13.4	-0.6
1996	10.6	23.5	14.1	0.5
1997	12.1	24	14	1.9
1998	13.4	23.1	15.4	1.6
1999	12.1	23.3	14.7	0.3
2000	11.6	24.2	13.7	0.3
2001	12.2	24.2	14.3	1.5
2002	12.7	23.1	12.5	0.9
2003	11.9	22.3	14.7	1.4
2004	12.2	24	14.7	0.5
2005	11.7	24.1	14.8	-0.1
2006	11.5	23.6	15.1	2.4
2007	12.1	23.8	14.5	0.7
2008	12.5	23.7	15.1	1.7
2009	12.6	23.3	14.7	0.5
2010	10.8	24.9	14.5	-0.7
2011	11	24	15.3	-0.4

Practice 2.4 Solution

- If you select 'Line Graph' from the 『eStatH』 menu using the QR on the left, the data input window as shown in <Figure 2.14> appears.
- Enter the 'main title', 'y title', and 'x title' as shown in the figure, and enter the spring, summer, autumn, and winter temperatures from 'Line 1' to 'Line 4'.Enter each variable name as Spring, Summer, Fall, Winter, and enter year data in 'X'.
- If you click the [Execute] button, a line graph as shown in <Figure 2.15> appears. You can also draw a line graph again by selecting the color of eacl line.



<Figure 2.14> Data input of average temperature by season for a line graph



<Figure 2.15> Line graph of average temperature by season in Korea

- Looking at the line graph of the average temperature by season, it can be seen that the average temperature is gradually increasing, especially in winter.

Exercise 2.3

The following table shows the average life expectancy of Koreans surveyed every 10 years from 1970 to 2020. Draw a line graph using 『eStatH』 to find out the characteristics.

[Table 2.7] Average life expectancy of Koreans

Year	Lif expectancy
1970	62.3
1980	66.1
1990	71.7
2000	76.0
2010	80.2
2020	81.3

Exercise 2.4

The table below shows the amount of imports and exports of Korea from 2001 to 2020. Draw a line graph using 『eStatH』 to find out the characteristics.

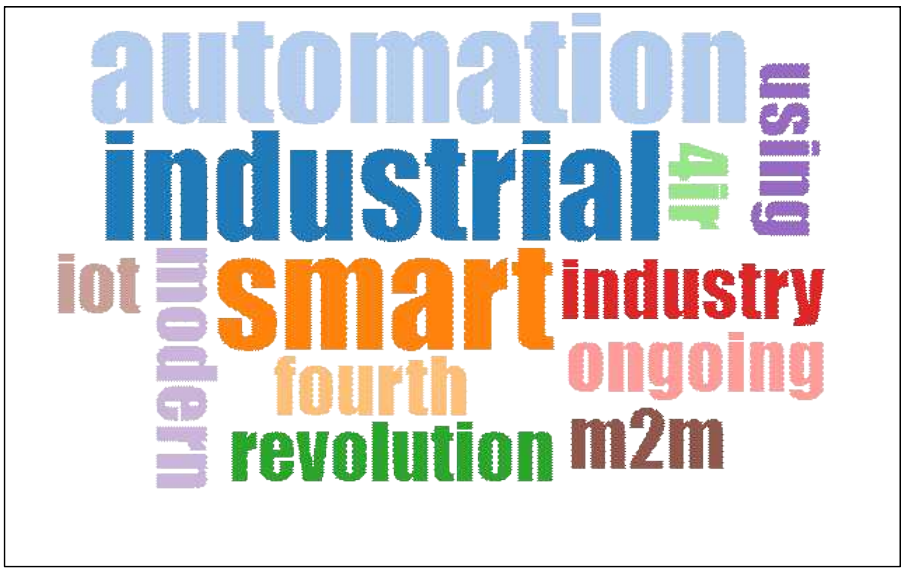
[Table 2.8] Amount of imports and exports of Korea
(unit: billon dollar)

Year	Export	Import
2001	150.4	141.1
2002	162.5	152.1
2003	193.8	178.8
2004	253.8	224.5
2005	284.4	261.2
2006	325.5	309.4
2007	371.5	356.8
2008	422.0	435.3
2009	363.5	323.1
2010	466.4	425.2
2011	555.2	524.4
2012	547.9	519.6
2013	559.6	515.6
2014	572.7	525.5
2015	526.8	436.5
2016	495.4	406.2
2017	573.7	478.5
2018	604.9	535.2
2019	542.2	503.3
2020	512.5	467.6

2.3 Word Cloud

☞ Think	The following is Wikipedia's explanation of the 4th industrial revolution. [Table 2.9] Wikipedia's explanation of the 4th industrial revolution. The Fourth Industrial Revolution (4IR or Industry 4.0) is the ongoing automation of traditional manufacturing and industrial practices, using modern smart technology. Large-scale machine-to-machine communication (M2M) and the internet of things (IoT) are integrated for increased automation, improved communication and self-monitoring, and production of smart machines that can analyze and diagnose issues without the need for human intervention.[1]
Explore	What words are important to explain the 4th industrial revolution?

- A **word cloud** is a visual representation of information in text data. The word cloud examines the frequency of all the words appearing in a given document, and displays the importance of the frequently appearing words with font size, inclination, or color. <Figure 2.16> is a word cloud for the explanation of the 4th industrial revolution above.

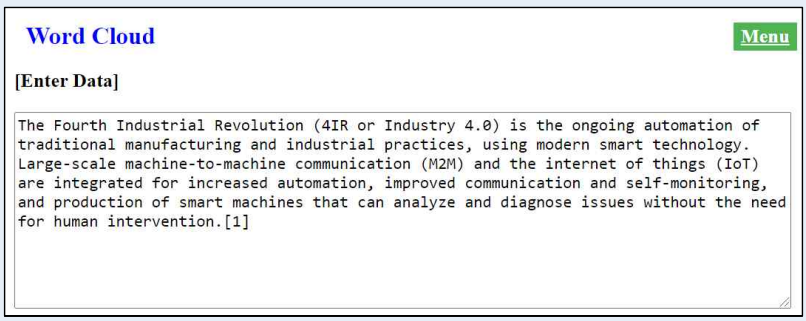


<Figure 2.16> Word cloud for Wikipedia's explation of the 4th industrial revolution

- The word cloud is useful for quickly recognizing the most prominent words in a docmuent and determining their relative importance. It can be used, for example, to visualize important topics in a political speech, or as a tool to determine hyperlinks to items related to a single word in social media software. Key words in the word cloud are also used as marketing terms related to a particular website.

- There are many algorithms for a word cloud generation, and 『eStat』 adopts the algorithm of d3 open software. d3 does not yet provide an accurate word cloud because there is no algorithm to remove unnecessary terms. Currently, the development of an algorithm that can display correlated words in a word cloud is in progress.

Practice 3.3	Draw and observe the word cloud for the 4th industrial revolution in [Table 2.9] using 『eStatH』.
Solution	• If you select 'Word Cloud' from the 『eStatH』 menu using the QR on the left, a window for data input as shown in <Figure 2.17> appears. • After copying the sentences in [Table 2.9], click the [Execute] button, and a word cloud as shown in <Figure 2.16> appears.



<Figure 2.17> Data input for Word Cloud

Exercise 2.3	The following are excerpts from US President John F. Kennedy's inaugural address. Create a word cloud for this inaugural address and analyze the key words.

[Table 2.11] Inaugral address of US President John F. Kennedy

In the long history of the world, only a few generations have been granted the role of defending freedom in its hour of maximum danger. I do not shrink from this responsibility--I welcome it. I do not believe that any of us would exchange places with any other people or any other generation. The energy, the faith, the devotion which we bring to this endeavor will light our country and all who serve it--and the glow from that fire can truly light the world.

And so, my fellow Americans: ask not what your country can do for you--ask what you can do for your country.

My fellow citizens of the world: ask not what America will do for you, but what together we can do for the freedom of man.