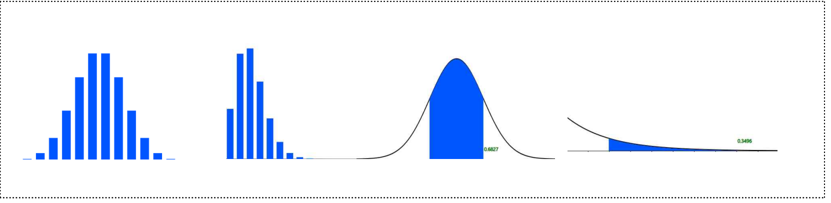
5

Probability Distribution



SECTIONS

CHAPTER OBJECTIVES

Similar events occur repeatedly in our everyday life. We describe a chance of the event using a concept of the probability.

In this chapter, we describe a definition of the probability in Section 5.1. Several rules of calculating the probability is described in Section 5.2.

There are many possible events in our everyday life which we can calculate the probability of the event. We focus on events which occur frequently and can be modeled as a random variable. Probability of all possible events of the random variable is called a probability distribution function.

In section 5.3, probability distributions of the discrete random variable such as Binomial Distribution, Poisson distribution, Geometric Distribution, Hypergeometric Distribution are described.

In section 5.4, probability distributions of continuous random variables such as Normal Distribution, and Exponential Distribution are described.

5.1 Definition of Probability

- Similar events occur repeatedly or are carried out in our everyday life. Let us consider following examples.

  - A machine produces products repeatedly at a production plant. The product can be either a normal product or a defective product, but it is not known in advance.
  - I have ordered pizza at home every Sunday. It usually takes about 30 minutes for a pizza to be delivered to the house, but the exact time is not known.

- These examples have common characteristics as follows:

  - There is repetition of similar events, such as the production of products or pizza deliveries.
  - We know all possibile outcomes of events that the product will be defective or normal, and the pizza delivery time will be 10 minute, 20 minutes, 30 minutes and so on.
  - But we do not know which outcome exactly will happen.

  Events with these three characteristics are subject to study and to apply statistics.

- An experiment in which a particular outcome occurs among known possible outcomes is called a **statistical experiment**. In a statistical experiment, the resulting outcome is uncertain and is determined by chance. For example, if you throw a coin, possible outcomes are either the head or tail, but the resulting outcome will come out by chance, so the experiment of tossing a coin is a statistical experiment. If a production plant produces products at one machine and their possible outcomes are either defective or normal, then the experiment of producing a product is a statistical experiment. Also, a pizza delivery time to your home which takes between 20 and 60 minutes is a statistical experiment.

Definition	**Statistical experiment** An experiment in which a particular outcome occurs among known possible outcomes. The outcome is uncertain and is determined by chance.

- In a statistical experiment, the set of all possible outcomes is referred to as a **sample space**, and a subset of this sample space is referred to as an **event**. The sample space is usually denoted by S, and events of the sample space are denoted by English capital letters as A, B, and C ... In the example above that a machine produces a product, the sample space is S = {normal, defective} and a subset of the sample space such as A = {defective} is an event. As such, when the number of elements in a sample space is either finite or countably infinite, it is called a **discrete sample space**. In a statistical experiment of the pizza delivery time to home, the sample space is all possible time between 20 minutes and 60 minutes, i.e. S = { (20,60) }. Delivery time between 20 minutes and 30 minutes ({20,30}) is called an event. As such, when the number of elements in a sample space is uncountably infinite, it is called a **continuous sample space**.

Definition	**Sample space and event**
	A set of all possible outcomes in a statistical experiment is referred to as a **sample space.** If the number of elements in a sample space is finite or countably infinite, it is called a **discrete sample space**. If the number of elements in a sample space is uncountably infinite, it is called a **continuous sample space**. A subset of the sample space is referred to as an **event**.

- A concept of probability is used to indicate the possibility of an event occurring in a statistical experiment. The **probability** is a representation of the likelihood of an event occurring using a number between 0 and 1. If an event is likely to occur, the probability is expressed as a number close to 1. If it is unlikely to occur, the probability is expressed as a number close to zero. Specifically, there are several ways to define the probability of an event using a number between zero and one. We introduce two definitions of the probability, one is a classical definition and the other is a relative definition of the probability.

Definition	**Classical definition of probability**
	Assume that all elements in sample space are likely to occur equally. The probability of an event A will occur, denoted as P(A), in case of the discrete sample space is defined as follows: $$P(A) = \frac{\text{Number of elements belonging to an event A}}{\text{Total number of elements in sample space}}$$ The probability of an event A will occur in case of the continuous sample space is defined as follows: $$P(A) = \frac{\text{Measurement of elements belonging to an event A}}{\text{Measurement of the total elements of a sample space}}$$ where measurement here can be the length, area and volume etc.

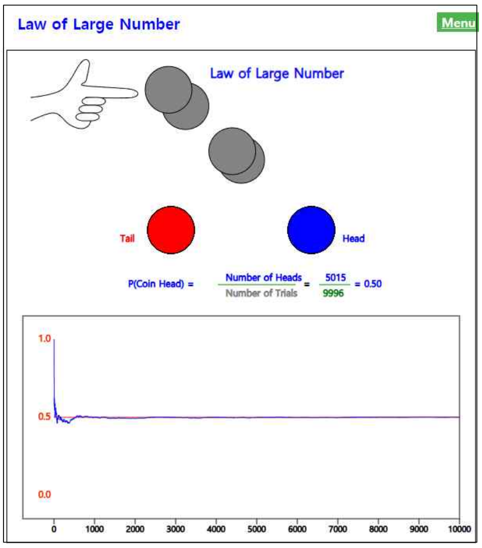
Example 5.1.1	An office worker went on a business trip to a city and there are two restaurants (Restaurant A and B) near his lodging. He was hesitating about which restaurant to go, and threw a dice to count the number of points that appear on the top. If he had odd numbers, he would go to the Restaurant A, and if he had even numbers, he would go to the Restaurant B. What is the probability that the Restaurant A would be picked?
Answer	◆ The sample space in this statistical experiment, which counts the number of points on the top by throwing a dice, is {1, 2, 3, 4, 5, 6}, and the number of odd events is {1, 3, 5}, so there are three elements. Therefore, the probability that restaurant A will be selected is 3/6 = 1/2.

Example 5.1.2	I ordered a pizza from home every Sunday. The time it takes for a pizza to be delivered my home has the same possibility for any time from 10 to 30 minutes (you may have a decimal number). What is the probability that a pizza will be delivered between 20 and 25 minutes?
Answer	◆ The sample space in this example is all values from 10 to 30 minutes collectively { (10,30) }, and where a pizza is delivered between 20 and 25 minutes is the event { (20,25) }. Therefore, the probability of this event is (25-20) / (30-10) = 0.25 by measuring the distance of the interval.

- The classical definition of the probability does not usually have a major problem in calculating the probability of a real problem. However, in the classical definition of the probability, it may not be possible to assume that all elements of the sample space are likely to occur. For example, tossing a coin usually results in ‘Head' and 'Tail' but can be very rarely ’Edge’. Considering this, the assumption that the sample space {'Head’, 'Tail' and 'Edge'} does not make sense that at this time each element of the sample space is likely to occur equally. The relative frequency definition of the probability is to solve these problems.

Definition	**Relative frequency definition of probability** The probability that an event A will occur, denoted by P(A), is the rate at which the event A occurs when many statistical experiments are conducted under the same condition repeatedly.

- If this definition is used, it can be explained that the coin tossing stands as an 'Edge.' If a coin has been thrown 10,000 times and 'Head' is appeared 4980 times, 'Tail' 5018 times, 'Edge' twice, then P({Head}) = 4980/10000, P({Tail}) = 5018/10000, P({Edge}) = 2/10000. More iterative runs make the definition of the probability using the relative frequency almost approximates to the probability values by the classical definition.
- <Figure 5.1.1> shows a simulation of a coin toss experiment using 『eStatU』 which shows the probability of ‘Head’ occurrence converges to one-half. This convergence of the probability is called the **law of large numbers**.



<Figure 5.1.1> Law of large number simulation

5.2 Calculation of Probability

- In order to calculate the probability of an event in a discrete sample space, the number of elements in the sample space and the number of elements included in the event should be counted. If all possible outcomes of the sample space is not large, the probability can be simply calculated, but it is generally not easy to count the number of possible outcomes. Effective methods for counting the number of complex cases include the permutation and combination.

Definition

Permutation

The number of cases to select r objects out of n objects considering the order is called the permutation and is calculated as follows:

$$_nP_r \ = \ n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

Therefore, the number of cases to list all n objects is as follows:

$$_nP_n \ = \ n(n-1)(n-2)\cdots 2\cdot 1 \ = \ n!$$

Note: $0! = 1$

Definition

Combination

The number of cases to select r objects out of n objects without considering the order is called the combination and is calculated as follows:

$$_nC_r \ = \ \frac{_nP_r}{r!} \ = \ \frac{n!}{r!(n-r)!}$$

Example 5.2.1	Four people A, B, C and D are intended to be placed on four side-by-side chairs. Calculate the total number of cases in which four people are placed, and the number of cases in which A is placed on the leftmost. What is the probability that A is placed on the leftmost side?
Answer	◆ The number of elements in the sample space in this example is as follows: (Number of people that can be placed on the leftmost) × (number of people except left who can be placed in the second position) × (number of people who can be placed in the third place except for two left people) × (number of people, excluding the three on the left, who can be positioned to the right) = 4 × 3 × 2 × 1 = 4! = 24 ◆ The event in which A is placed on the left is the number of people placed in the second, third, and right positions except A, so 3×2×1 = 3!. Therefore, the probability that A will be placed to the left is as follows: 3! / 4! = 6 / 24 = 0.25

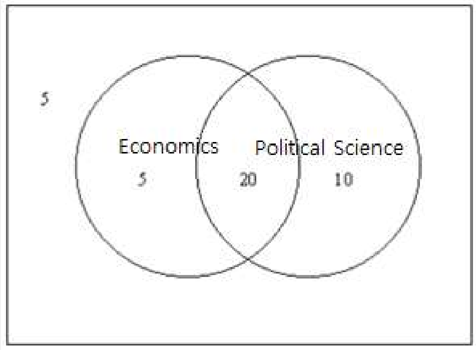
Example 5.2.2	A company has four security guards (A, B, C, D). Each morning, two of these guards are randomly selected, one at the front gate and the other at the rear guard. Obtain the total number of cases in which four people are placed at the front and rear gates and the number of cases in which A is placed at the front gate. What is the probability that A will be placed at the front gate?
Answer	◆ The number of elements in the sample space in this problem is as follows: (number of people who can be placed at the front gate) × (number of people who can be placed in the rear except those placed in the front) = 4 × 3 = ${}_4P_2$ = 12 ◆ The number of elements in the event where A will be placed at the front gate is ${}_3P_1$ = 3, since A can be placed at the front gate and one of the other three can be placed at the rear gate. That is, the probability that A will be placed at the front gate one day is as follows: $\frac{{}_3P_1}{{}_4P_2} = \frac{3\times 1}{4\times 3} = \frac{1}{4}$

- There are several calculation rules for calculating complex probabilities other than the permutation and the combination. Let us consider the examples below to explain the rules.

Example 5.2.3	Out of 40 sophomores in a Statistics Department this semester, 25 students are taking Economics, 30 students are taking Political Science and 20 students are taking both courses. When I meet one of the sophomores, what is the probability of this student taking either Economics or Political Science (that is one or both)?

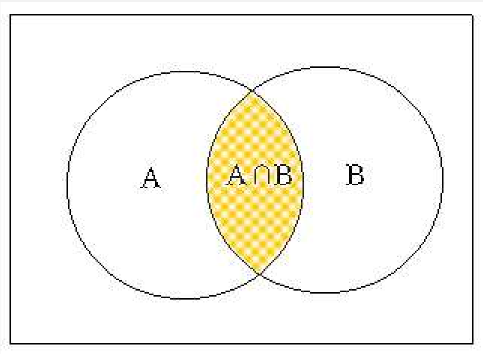
Example 5.2.3 Answer

- Since there are 25 students who take Economics and 20 students taking both courses, 25 - 20 = 5 students take only Economics. Also, since there are 30 students who take Political Science, 30 - 20 = 10 students take only Political Science. Thus, as shown in <Figure 5.2.1>, the number of students taking either Economics or Political Science is 5 + 10 + 20 = 35. Therefore, the probability of students taking either Economics or Political Science is 35 / 40.



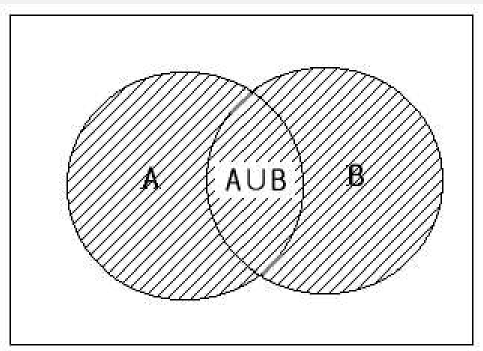
<Figure 5.2.1> Students who take either economics or political science

- Consider the case of students taking both Economics (A) and Political Science (B). The event that a student takes both courses are denoted as A ∩ B and is called an **intersection event** of A and B (<Figure 5.2.2>).



<Figure 5.2.2> Intersection events A ∩ B

- The event that a student takes either Economics or Political Science (one or both) is denoted as A ∪ B and is called an **union event** of A and B (<Figure 5.2.3>).



<Figure 5.2.3> Union Event A ∪ B

Example 5.2.3 Answer (continued)	◆ Probabilities of these events on this example are as follows: $P(A) = 25 / 40$ $P(B) = 30 / 40$ $P(A \cap B) = 20 / 40$ $P(A \cup B) = 35 / 40$ ◆ The probability of $P(A \cup B)$ can also be calculated as follows if you look at the <Figure 5.2.1>. $P(A \cup B) = P(A) + P(B) - P(A \cup B)$ $= 25/40 + 30/40 - 20/40 = 35/40$ ◆ That is, the probability of taking either Economics or Political Science, $P(A \cup B)$, can be calculated by adding the probability of taking each course and then by subtracting the probability of taking both courses.

- The rule discussed on [Example 5.2.3] is called the **addition rule of probability.**

Definition	**Addition Rule of Probability** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ If $A \cap B = \varnothing$, then the rule becomes as follows: $P(A \cup B) = P(A) + P(B)$ In this case, the events A and B are called **mutually exclusive events**.

Example 5.2.4	In [Example 5.2.3], if there are 10 students taking Economics, 20 students taking Political Science, and if there are no students taking both courses, what is the probability of a student is taking either Economics or Political Science?
Answer	◆ In this case, because there are no students taking both courses, the events in which they take Economics (A) and Political Science (B) are mutually exclusive. Thus, the probability to take either Economics or Political Science, $P(A \cup B)$, is as follows: $P(A \cup B) = P(A) + P(B) = 10/40 + 20/40 = 0.75$

- Let us consider the example below to find out the multiplication rule of probability.

Example 5.2.5

Students of ADA University come from either Baku or a province. Among the 30 sophomores in the Department of Economics, there are 10 males and 20 females, one of males and five of females are from the province.

1) When selecting a student, what is the probability that he is from the province?
2) When I selected a student, she was a female. What is the probability that this student is from the province?
3) When I selected a student, he was from the province. What is the probability of this student being a male?
4) When selecting a student, what is the probability that he is a male and from Baku?

Answer

- To solve this problem, it is convenient to organize the information given into a cross table as shown below.

	Baku	Province	Total
Male	_____	1	10
Female	_____	5	20
Total	_____	_____	30

- If we calculate and insert the blanks on the above table, it is as follows. Let us call the event of male as M, the female as F, from Baku as B, from the province as C.

	Baku(B)	Province(C)	Total
Male(M)	9	1	10
Female(F)	15	5	20
Total	24	6	30

1) $P(C) = 6/30$.

2) The probability that this student is from the province among females is 5/20. This probability is denoted as $P(C \mid F)$ and is called a **conditional probability**.

3) The probability of a male from the province is $P(M \mid C) = 1/6$.

4) The probability is $P(M \cap B)$ and the cross table shows that the answer is 9/30. Alternatively, the probability of being a male among all students can be first obtained as $P(M) = (10/30)$ and then multiplied by the conditional probability of being from Baku among males, $P(B \mid M)) = 9/10$. Namely

$$P(M \cap B) = P(M)\ P(B \mid M) = (10/30) \times (9/10) = 9/30$$

This expression shows that the conditional probability $P(B \mid M)$ can be calculated by dividing $P(M \cap B)$ by $P(M)$.

$$P(B \mid M) = \frac{P(M \cap B)}{P(M)} = \frac{9/30}{10/30} = \frac{9}{10}$$

- In addition, the probability $P(M \cap B)$ can be obtained first by the probability of being a student from Baku, $P(B) = 24/30$, and then multiplied by the probability of being a male from Baku ($P(M \mid B) = 9/24$).

$$P(M \cap B) = P(B)\ P(M \mid B) = (24/30) \times (9/24)$$

- Conditional probability is generally defined as follows:

Definition

Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{if} \quad P(B) \neq 0$$

- In the above example, the probability of an intersection event is expressed by multiplying the probabilities of other events and it is called the **multiplication rule of probability**.

Definition

Multiplication Rule of Probability

$$P(A \cap B) = P(A)\ P(B \mid A)$$

If $P(B \mid A) = P(B)$, then the rule becomes as follows:

$$P(A \cap B) = P(A)\ P(B)$$

In this case, the events A and B are called **independent events**.

Example 5.2.6 Tiger team of professional baseball has the probability of 0.7 to beat Lion team recently. What is the probability that Tiger is winning both games in this evening's double match? Assume that winning one game does not affect winning the next.

Answer

- Let us call the event that the Tiger wins the first game is A and the event that the Tiger wins the second game is B. Since A and B are independent of each other, the probability that the Tiger is winning both games is as follows:

$$P(A \cap B) = P(A)\ P(B) = 0.7 \times 0.7 = 0.49$$

Example 5.2.7 The following is a table of 30 second-year students by gender and region of origin. Are the events of male and Baku origin independent of each other?

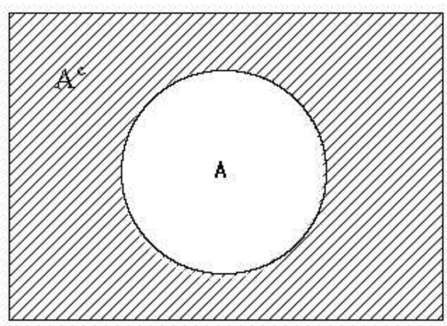
	Baku(B)	Province(C)	Total
Male(M)	5	5	10
Female(F)	10	10	20
Total	15	15	30

Example 5.2.7 Answer	◆ Let us call the event of male as M, female as F, from Baku as B, and from province as C. From the table, probabilities of P(M ∩ B), P(M) and P(B) are as follows: P(M ∩ B) = 5/30 P(M) = 10/30 P(B) = 15/30 ◆ Therefore, the following relationship is satisfied: P(M ∩ B) = P(M) P(B) ◆ The events of male and Baku origin are independent of each other. Note that P(M ∣ B) = 5/15 = 1/3 P(M)=10/30 so, P(M ∣ B) = P(M). In this case, all events of both M and C, F and B, F and C are independent of each other. We call the two attributes, gender and region are independent of each other. In [Example 5.2.5], gender and region are not independent of each other.

- The following is an example of how to calculate the probability of a complementary event.

Example 5.2.8	There is a box of six products, two of which are defective. What is the probability that at least one defective product will be found when three have been extracted for product testing? Assume that the product is extracted once for inspection without replacement.
Answer	◆ The probability of finding one defective in the three product tests is as follows: $({}_4C_2 \times {}_2C_1) \ / \ {}_6C_3$ = 3/5. ◆ The probability of finding two defective products is as follows: $({}_4C_1 \times {}_2C_2) \ / \ {}_6C_3$ = 4/20 = 1/5. ◆ Thus, the probability that at least one defect will be found is 3/5 + 1/5 = 4/5. ◆ Another way to calculate this probability is to obtain the probability of an event in which there will be no defect (this is called a **complementary event**) and then, subtract it from 1. In other words, the probability that at least one defective product can be calculated as follows: $1 - ({}_4C_3 \ / \ {}_6C_3)$ = 1 - (4/20) = 4/5

- The method used in the above example is called a rule of probability calculation using a complementary event, and is often used to obtain the probability that the word 'at least' is contained. <Figure 5.2.4> is a picture of an event.



<Figure 5.2.4> Complementary event A^C

Definition	**Probability calculation using a complementary event** if A^C denotes a complementary event of the event A, then P(A^C) can be calculated as follows: P(A^C) = 1 - P(A)

5.3 Discrete Random Variable

- In case of statistical experiments which are frequently observed around us, there are many similar probability calculations. For example, the problem of tossing coins several times to see how many times the head comes out is similar to counting how many defective products are made from a product line. This problem is also similar to counting the number of voters who support a particular candidate for the president. In this section, the probability calculations as the previous examples, in general the discrete sample spaces are discussed.
- Consider a statistical experiment in which a coin is thrown repeatedly two times. If the coin is ideal, the sample space for this experiment is {'Tail-Tail', 'Tail-Head', 'Head-Tail', 'Head-Head'}. The probability of an event in which each element of the sample space is produced is 1/4 by the classical definition. In most cases, the fact that we are interested in this example will be counting the number of heads or tails. If X is defined as 'the number of heads' in this experiment, the possible value of X can be 0, 1, or 2 and we are interested in calculating probabilities that X=0, X=1 or X=2. As such, a function that corresponds to one real number between [0,1] for each element of the sample space is called a **random variable** (see Table 5.3.1).

Table 5.3.1 Random variable X = 'Number of Heads' when tossing a coin twice'

Sample space	X = 'Number of Heads'
Tail-Tail	0
Tail-Head	1
Head-Tail	1
Head-Head	2

- When possible values of a random variable are finite or countably infinite, it is called a **discrete random variable**. If possible values of a random variable are uncountably infinite, it is called a **continuous random variable** and discussed in more detail in Section 5.4.

Definition

Random Variable

Random variable is a function from the sample space to a real number in [0, 1]. When possible values of a random variable are finite or countably infinite, it is called a **discrete random variable**. If possible values of a random variable are uncountably infinite, it is called a **continuous random variable**.

- The probability that the random variable X defined as in Table 5.3.1 will be zero denoted as P(X=0) is 1/4, because it is the probability of an event {Tail-Tail}. The probability that X being 1, P(X=1), is 2/4, because, P({Tail-Head, Head-Tail}) is 2/4. Also, the probability that X being 2, P(X=2), is 1/4 because P({Head-Head}) is 1/4. The probabilities for each value of the random variable X can be summarized as shown in Table 5.3.2, and it is called a **probability distribution function** of X usually denoted as f(x). <Figure 5.3.1> is a graph of f(x).

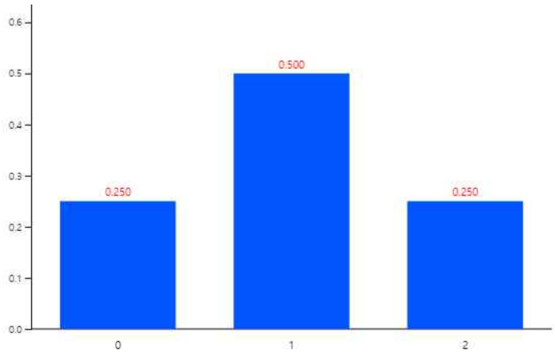
Table 5.3.2 Probability distribution function of X = 'Number of Heads' when tossing a coin twice'

1) Table style of the probability distribution function

X = x	P(X=x)
0	1/4
1	2/4
2	1/4
Total	1

2) Function style of the probability distribution function

$$
\begin{aligned}
f(x) &= 1/4, \quad x=0 \\
&= 2/4, \quad x=1 \\
&= 1/4, \quad x=2
\end{aligned}
$$



<Figure 5.3.1> Probability distribution function of the random variable X = 'Number of Heads' when tossing a coin twice'

- The cumulative probability of P(X $\leq$ x) as the value of random variable X increases is referred to as a **cumulative distribution function** and denoted as F(x). In the previous example, the cumulative distribution function of the random variable X = 'Number of Heads' when tossing a coin twice' is shown as in Table 5.3.3.

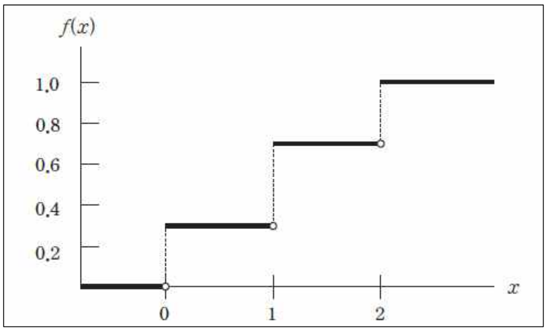
Table 5.3.3 Cumulative distribution function of the random variable X = 'Number of Heads' when tossing a coin twice'

1) Table style of the cumulative distribution function

X = x	P(X≤x)
0	1/4
1	3/4
2	4/4=1

2) Function style of the cumulative distribution function

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/4, & 0 \le x < 1 \\ 3/4, & 1 \le x < 2 \\ 1, & 2 \le x \end{cases}$$



<Figure 5.3.2> Cumulative distribution function of the random variable X = 'Number of Heads' when tossing a coin twice'

Definition

If probability for each value of the random variable X is summarized as a function, it is called a **probability distribution function** of X and usually denoted as f(x).

The cumulative probability of P(X ≤ x) as the value of random variable X increases is referred to as a **cumulative distribution function** and denoted as F(x).

Example 5.3.1

There are 200 families living in a village. The number of visits to hospitals by each household over the past year is as follows. Obtain the probability distribution function and the cumulative distribution function of the random variable X = 'hospital visit count'.

Hospital visit	0	1	2	3	4
Household	74	80	30	10	6

Answer

Probability distribution function

X = x	P(X=x)
0	0.37
1	0.40
2	0.15
3	0.05
4	0.03
Total	1.00

Cumulative distribution function

X = x	P(X≤x)
0	0.37
1	0.77
2	0.92
3	0.97
4	1.00

- If possible values of a discrete random variable X are $x_1, x_2, \cdots, x_n$, a mean and variance of X are also used as measures of the central tendency and dispersion. The mean of X called an **expectation** of X, denoted E(X) or μ, and the variance of X, denoted as V(X) or σ^2, are defined as follows. The standard deviation of X, denoted σ, is the square root of the variance X.

Definition	Expectation and Variance of X
	$E(X) \;=\; \mu \;=\; \sum_{i=1}^{n} x_i \, P(X=x_i)$ $V(X) \;=\; \sigma^2 \;=\; \sum_{i=1}^{n} (x_i-\mu)^2 \, P(X=x_i) \;=\; \sum_{i=1}^{n} x_i^2 \, P(X=x_i) \,-\, \mu^2$

Example 5.3.2	Find the expected value and variance of the random variable X = 'Number of Heads' when tossing a coin twice' which described in Table 5.3.2.
Answer	◆ Expectation and variance of X are as follows: $\mathrm{E(X)} \;=\; \mu \;=\; \sum_{i=1}^{n} x_i \mathrm{P}(\mathrm{X}=x_i) \;=\; 0\times\frac{1}{4}+1\times\frac{2}{4}+2\times\frac{1}{4} \;=\; 1$ $\mathrm{V(X)} \;=\; \sum_{i=1}^{n} x_i^2 \mathrm{P}(\mathrm{X}=x_i)-\mu^2 \;=\; 0^2\times\frac{1}{4}+1^2\times\frac{2}{4}+2^2\times\frac{1}{4}-1^2 \;=\; \frac{1}{2}$

- When knowing the expected value E(X) and variance V(X) of a random variable X, it is often necessary to obtain the expected value and variance of $a\mathrm{X}+b$ where a and b are constants. The expected value and variance of the new random variable $a\mathrm{X}+b$ are as follows. This formula applies equally to a continuous random variable.

Definition	Expectation and variance of $aX+b$ where a, b are constant.
	$\mathrm{E}(a\mathrm{X}+b) \;=\; a\mathrm{E(X)}+b$ $\mathrm{V}(a\mathrm{X}+b) \;=\; a^2\mathrm{V(X)}$

Example 5.3.3	The mean of midterm exam scores in a Statistics course was 60 points and the variance was 100. In order to adjust the scores, professor is thinking of following alternatives. Find the mean and variance of each alternative. 1) Add 20 points to each student's score. 2) Each student's score is multiplied by 1.4. 3) Multiply each student's score by 1.2 and add 10 points.
Answer	◆ The random variable X is the mid-term score and its mean and variance are E(X) = 60 and V(X) = 100.

Example 5.3.3 Answer (continued)	1) The mean and variance of the new random variable X + 20 are as follows: E(X + 20) = E(X) + 20 = 60 + 20 V(X + 20) = V(X) = 100 In other words, if you add 20 points to each score, the mean is increased 20 points but there is no change on the variance. 2) The mean and variance of the new random variable 1.4X are as follows: E(1.4X) = 1.4 E(X) = 1.4 × 60 = 84 V(1.4X) = 1.4^2 V(X) = 1.96 × 100 = 196 In other words, if you multiply 1.4 to each score, the mean is increased 1.4 times and the variance is increased 1.4^2 = 1.96 times. 3) The mean and variance of the new random variable 1.2X + 10 are as follows: E(1.2X + 10) = 1.2 E(X) + 10 = 1.2 × 60 + 10 = 82 V(1.2X + 10) = 1.2^2 V(X) = 1.44 × 100 = 144 In other words, adding 10 points after multiplying by 1.2 increases the mean by 1.2 times and the variance by 1.2^2 = 1.44 times. Note that adding scores will change the mean, but not the variance.

- Let the mean of random variable X be μ and the standard deviation σ. Then $Z = (X-\mu)/\sigma$ is a new random variable with the mean of 0 and the variance of 1. This new random variable is referred to as a **standardized random variable**.

Definition	**Standardized random variable** If the mean of a random variable X is μ, and the standard deviation is σ, then $Z = (X-\mu)/\sigma$ is a new random variable with the mean of 0 and the variance of 1. This new random variable is referred to as a standardized random variable.

- Let us discuss about binomial, Poisson, geometrical, and hyper-geometrical distributions which are used widely as discrete probability distributions.

5.3.1 Binomial Distribution

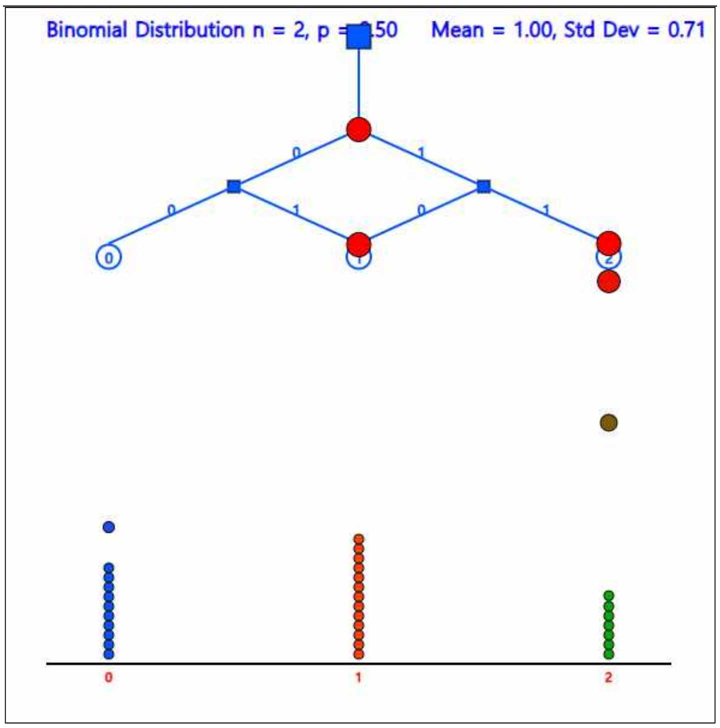
- Examples that are similar to experiments which examine how many times the head comes out by tossing coins are observed around us. Let us take a look at following examples.

  - Products produced in a factory machine are inspected and classified as either defective or normal.
  - In an election survey, ask voters whether he would vote for a candidate (pro) or not (con).

- In these examples there are two possible outcomes such as {defective, normal} and {pro, con} in each experiment and this experiment is repeated, but we do not know what the outcome of each experiment. Each experiment which has two possible outcomes is specifically called a **Bernoulli trial**, and one outcome of interest in the two is often referred to as 'success' and the other outcome as 'failure'. This Bernoulli trial is usually repeated several times and the number of ‘success’ is counted as following.

  - Throw a coin five times and examine the number of heads.
  - Inspect 100 products produced in a factory and count the number of defective products.
  - Count the number of voters in favor of a particular candidate for the president among 50 eligible voters.

- Counting the number of success on repeated Bernoulli trials can be simulated using the module of ‘Binomial Experiment’ in 『eStatU』. In this module, a ball is dropped from the top and if it hits a blue box, it has one-half chance to fall to the left (get zero point) or right (get one point). The dropped ball again falls to the left and right with a 1/2 chance such as <Figure 5.3.3>. The same experiment is repeated 100 times (drop 100 balls), and we examine the number of balls which got points 0, 1, and 2. It is similar to tossing a coin two times and counting the number of heads.



<Figure 5.3.3> Simulation of Binomial distribution

- The 'counting of success' when performing the independently repeated Bernoulli trial with the same probability of success is called a **binary random variable**, and its distribution is called a **binomial distribution**. The probability calculation of the binomial distribution will be found by the example below.

Example 5.3.4

Four more games will be played by the Tiger baseball team this season. If the Tiger team has a 60% chance of winning every game, what is the probability of the followings.

1) losing all of them?
2) winning only once?
3) winning twice?
4) winning three times?
5) winning all four times?
6) Find the probability distribution function of the random variable X = 'the number of games the tiger wins'.

Answer

- This problem is the enforcement of Bernoulli trial in each game of 'win' and 'fail'. This Bernoulli trial is repeated four times. The sample space is all about winning or losing game and there are $2^4 = 16$ elements shown as follows by marking the winning in O and the losing in X.

S = {'XXXX', 'OXXX', 'XOXX', 'XXOX', 'XXXO', 'OOXX', 'OXOX', 'OXXO', 'XOOX', 'XOXO', 'XXOO', 'OOOX', 'OOXO', 'OXOO', 'XOOO', 'OOOO'}

1) The event that the Tiger will loose all games is {'XXXX'} and the probability of this event is (0.4)×(0.4)×(0.4)×(0.4) = $(0.4)^4$.

2) There are four events that the Tiger is winning once and losing three times such as {'OXXX', 'XOXX', 'XXOX', 'XXXO'}. These four cases are equal to the number of O's in a single seat when there are four seats which is ${}_4C_1$. Since the probability of each event is (0.6)×(0.4)×(0.4)×(0.4), the probability of the Tiger winning once is ${}_4C_1(0.6)(0.4)^3$.

3) There are six events that the Tiger is winning two times and losing two times such as {'OOXX', 'OXOX', 'OXXO', 'XOOX', 'XOXO', 'XXOO'}. These six cases are equal to the number of O's in two seats when there are four seats which is ${}_4C_2$. Since the probability of each event is (0.6)×(0.6)×(0.4)×(0.4), the probability of the Tiger winning twice is ${}_4C_2(0.6)^2(0.4)^2$.

4) There are four events that the Tiger is winning three times and losing one time such as {'OOOX', 'OOXO', 'OXOO', 'XOOO'}. These four cases are equal to the number of O's in three seats when there are four seats which is ${}_4C_3$. Since the probability of each event is (0.6)×(0.6)×(0.6)×(0.4), the probability of the Tiger winning three times is ${}_4C_3(0.6)^3(0.4)^1$.

5) There is one event that the Tiger is winning four times such as {'OOOO'}. This one case is equal to the number of O's in four seats when there are four seats which is ${}_4C_4$. Since the probability of each event is (0.6)×(0.6)×(0.6)×(0.6), the probability of the Tiger winning all four times is ${}_4C_4(0.6)^4$.

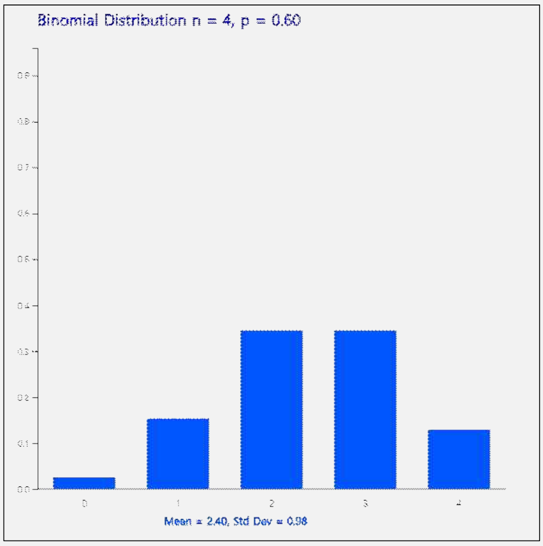
6) The probability distribution function of the random variable X = 'the number of games the Tiger wins' is a summary of the above probabilities.

x	P(X=x)
0	${}_4C_0(0.4)^4$ = 0.0256
1	${}_4C_1(0.6)(0.4)^3$ = 0.1536
2	${}_4C_2(0.6)^2(0.4)^2$ = 0.3456
3	${}_4C_3(0.6)^3(0.4)$ = 0.3456
4	${}_4C_4(0.6)^4$ = 0.1296

Example 5.3.5 By using『eStatU』, find the probability and the probability distribution function of [Example 5.3.4].

Answer

- Select 'Binomial Distribution' from the menu of『eStatU』and enter n = 4, p = 0.6 and press the [Execute] button to display a binomial function graph as shown in <Figure 5.3.4>. Table 5.3.4 shows the table when you click the [Binomial Prob Table] button. This table makes it easy to obtain Binomial distribution probabilities from [Example 5.3.4].



<Figure 5.3.4>『eStatU』Binomial distribution graph when n = 4, p = 0.6

Table 5.3.4『eStatU』Binomial distribution table when n = 4, p = 0.6

n = 4	p = 0.600		
x	P(X = x)	P(X ≤ x)	P(X ≥ x)
0	0.0256	0.0256	1.0000
1	0.1536	0.1792	0.9744
2	0.3456	0.5248	0.8208
3	0.3456	0.8704	0.4752
4	0.1296	1.0000	0.1296

- In general, the probability of 'success' when a Bernoulli trial is repeated n times, i.e., the probability of the binomial distribution, is as the following definition.

Definition

Binomial Distribution

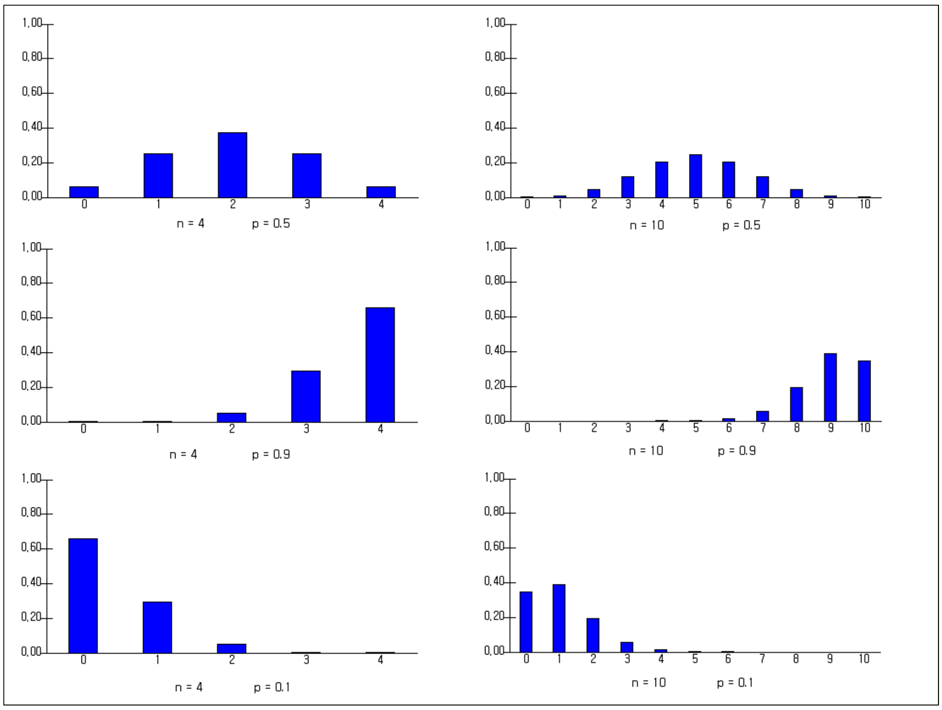
If the probability of success is p in a Bernoulli trial and the trial is repeated n times independently, the probability distribution function that the random variable X = 'the number of success' is x is as follows. It is called a **binomial distribution** and denoted as $B(n,p)$.

$$f(x) \;=\; {}_nC_x \; p^x \; (1-p)^{n-x} \quad , \qquad x = 0,1,2,...,n$$

The expectation and variance of the binomial distribution are as follows:

E(X) = np, V(X) = $np(1-p)$.

- In the binomial distribution function, n (number of trials) and p (success probability) are called parameters. <Figure 5.3.5> shows a binomial distribution of different n and p.



<Figure 5.3.5> Binomial distribution for various n and p

Example 5.3.6

Past experience shows that a salesperson from an insurance company has a 20% chance of insuring a customer when he meets. The salesperson is scheduled to meet 10 customers this morning. Calculate the following probabilities directly and check using 『eStatU』.

1) What is the probability that three customers will get insurance?
2) What is the probability that two or more customers will get insurance?
3) How many people on average would sign up? And what is its standard deviation?

Answer

- This is a Binomial distribution with n = 10, p = 0.2.

1) The probability that three customers will get insurance is as follows:

P(X=3) = ${}_{10}C_3 (0.2)^3 (1-0.2)^{10-3}$ = 0.2013

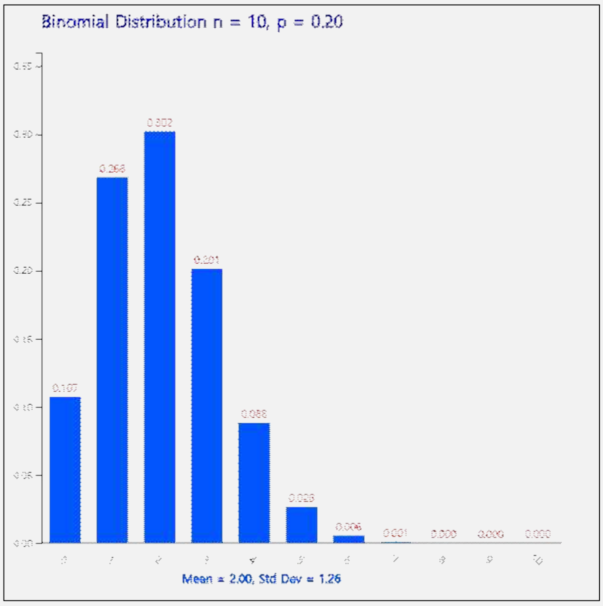
2) The probability that two or more customers will get insurance may use the complement event as follows:

P(X ≥ 2) = 1 - P(X=0) - P(X=1)
= 1 - ${}_{10}C_0 (0.2)^0 (1-0.2)^{10}$ - ${}_{10}C_1 (0.2)^1 (1-0.2)^{10-1}$
= 1 - 0.1074 - 0.2684 = 0.6242

3) Expectation and standard deviation are as follows:

E(X) = np = 10 × 0.2 = 2
V(X) = np(1-p) = 10 × 0.2 × 0.8 = 1.6
Standard deviation = $\sqrt{1.6}$ = 1.265

- Select ‘Binomial Distribution’ from the menu of 『eStatU』, enter n=10, p=0.2, and click on the [Execute] button to display the graph shown in <Figure 5.3.6>. Checking 'Show Probability' option shows the probability on each bar where you can see the values in the above calculations.



<Figure 5.3.6> 『eStatU』 Binomial distribution when n = 10, p = 0.2

Example 5.3.6 Answer (continued)

◆ Pressing the [Binary Prob Table] button will show the Binomial distribution table shown in Table 5.3.5. From here you can see that P(X ≥ 2) = 0.6242

Table 5.3.5 『eStatU』 Binomial Distribution Table when n = 10, p = 0.2

n = 10	p = 0.200		
x	P(X = x)	P(X ≤ x)	P(X ≥ x)
0	0.1074	0.1074	1.0000
1	0.2684	0.3758	0.8926
2	0.3020	0.6778	0.6242
3	0.2013	0.8791	0.3222
4	0.0881	0.9672	0.1209
5	0.0264	0.9936	0.0328
6	0.0055	0.9991	0.0064
7	0.0008	0.9999	0.0009
8	0.0001	1.0000	0.0001
9	0.0000	1.0000	0.0000
10	0.0000	1.0000	0.0000

[Practice 5.3.1]

It is said that 60% of car drivers usually use a seat belt. When you select 15 drivers randomly, find the following probabilities for the number of drivers who normally use the seat belt. Check the calculation using 『eStatU』.

1) Probability of 10 or more drivers.
2) Probability of 8 or less drivers.
3) Probability of at least 11 drivers.
4) Probability of at least 7 drivers.

- If the value of n increases, it is not easy to calculate the probability of Binomial probability distribution even with a calculator. In 『eStatU』, the probability of a case $n \le 100$ is readily available.

Example 5.3.7

The defect rate of electronic parts produced in a factory is 5 percent. When you have a box containing 50 of these parts, use 『eStatU』 to obtain the following probabilities.

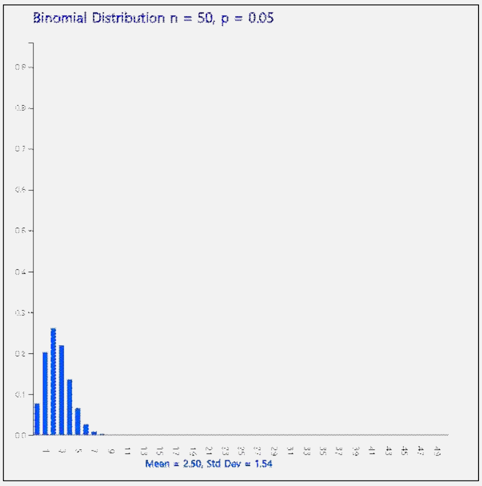
1) What is the probability of having no defective product?
2) What is the probability of having 1 to 3 defective products?
3) What is the probability of having more than three defective products?

Answer

◆ When you select n=50, p=0.05 from the 'Binomial Distribution' of 『eStatU』 and click on the [Execute] button, the graph such as <Figure 5.3.7> appears. If you click the [Binomial Prob Table] button, then Table 5.3.6 appears.

Example 5.3.7 Answer (continued)

1) You can check P(X=0) = 0.0769 easily from the table.



<Figure 5.3.7> 『eStatU』 Binomial Distribution when n = 50, p = 0.05

Table 5.3.6 『eStatU』 Binomial Distribution Table when n = 50, p = 0.05 (in part)

n = 50	p = 0.050		
x	P(X = x)	P(X ≤ x)	P(X ≥ x)
0	0.0769	0.0769	1.0000
1	0.2025	0.2794	0.9231
2	0.2611	0.5405	0.7206
3	0.2199	0.7604	0.4595
4	0.1360	0.8964	0.2396
⋯	⋯	⋯	⋯

2) The probability of having 1 to 3 defective products is $P(1 \le X \le 3)$ and it can be calculated as follows:

$$P(1 \le X \le 3) = P(X \le 3) - P(X \le 0) = 0.7604 - 0.0769 = 0.6835$$

You may calculate this probability as P(X=1) + P(X=2) + P(X=3).

3) The probability of having more than three defective products can be calculated by using the Table 5.3.6 as $P(X \ge 3) = 0.4595$. You may calculate this probability by using the complementary event as follows:

$$P(X \ge 3) = 1 - P(X \le 2) = 1 - 0.5405 = 0.4595$$

[Practice 5.3.2]

A salesperson found that there was a 30% chance of selling a product when a customer visited. If one day ten customers visit this salesman, calculate following probabilities using 『eStatU』.

1) Exactly how likely is it to sell three products?
2) What is the probability of selling three or more products?
3) What is the probability of selling less than 3 products?
4) What are the odds that none of them could be sold?
5) What is the probability of selling 5 products?

- If the number of trials n of the Binomial Distribution is greater than 100, the probability calculation can not be obtained even using 『eStatU』. In such cases, you can use a normal approximation with the mean np and variance $np(1-p)$ which is described in Section 5.4.2.

5.3.2 Poisson Distribution

- Consider the following examples that are frequently observed in many areas around us.

  - The number of calls made to an office of the Economics Department between 9 am and 10 am daily for one month.
  - The number of traffic accidents occurring at a certain intersection every day is investigated for one year.
  - The number of defective spots per each one square meter of the fabric is investigated for 100 square meters.
  - The number of typing errors that occur on each page of a book.
  - The number of accidents occurring during a week in a factory for one year

- What these statistical experiments have in common is to investigate the number of events per unit time or unit area. A random variable that represents this 'occurrence of events per unit time or unit area' is called a **Poisson random variable** and its distribution is called a **Poisson distribution**.
- Probability of the Poisson distribution can be calculated using the following formula.

Definition

Poisson Distribution

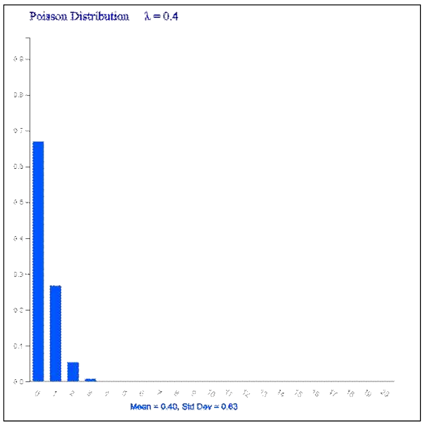
The distribution of a Poisson random variable X = 'Occurrence of success event per unit time or unit area' is as follows when the average number of success is λ.

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0,1,2, \ldots$$

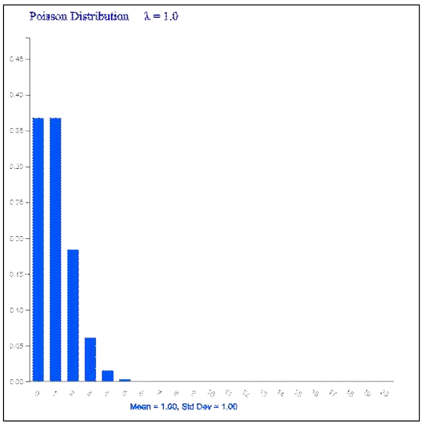
The expectation and variance of the Poisson random variable are as follows:

E(X) = λ
V(X) = λ

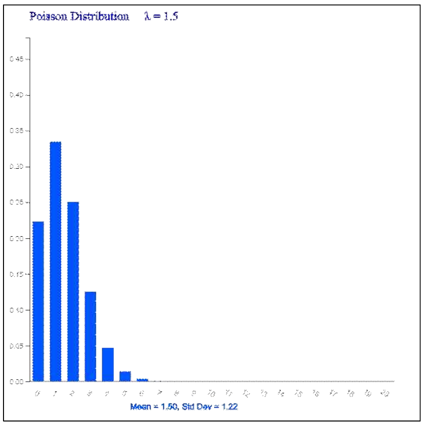
- The average number of success λ in the Poisson distribution function is called a parameter of the Poisson distribution. Note that the mean and variance of the Poisson distribution are the same as λ. <Figure 5.3.8> to <Figure 5.3.11> show the Poisson distributions for different values of λ by using 『eStatU』.



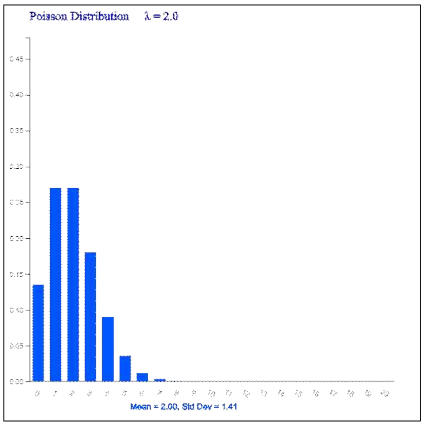
<Figure 5.3.8> Poisson Distribution when λ = 0.4



<Figure 5.3.9> Poisson Distribution when λ = 1.0



<Figure 5.3.10> Poisson Distribution when λ = 1.5



<Figure 5.3.11> Poisson Distribution when λ = 2.0

Example 5.3.8

Assume that cars arriving at a highway toll gate per one minute during rush hour is the Poisson distribution with an average of five cars. One day, if you observe the toll gate for one minute during rush hour, calculate the following probabilities.

1) What is the probability that none of cars will arrive?
2) What is the probability of five cars arriving?
3) What is the probability of more than two cars arriving?

Answer

- Let X be the Poisson random variable with λ = 5.

1) P(X = 0) = f(0) = $\frac{e^{-5}5^0}{0!}$ = 0.0067

2) P(X = 5) = f(5) = $\frac{e^{-5}5^5}{5!}$ = 0.1755

3) P(X $\geq$ 2) = 1 - P(X $\leq$ 1) = 1 - P(X=0) - P(X=1)
= 1 - 0.0067 - 0.0337 = 0.9596

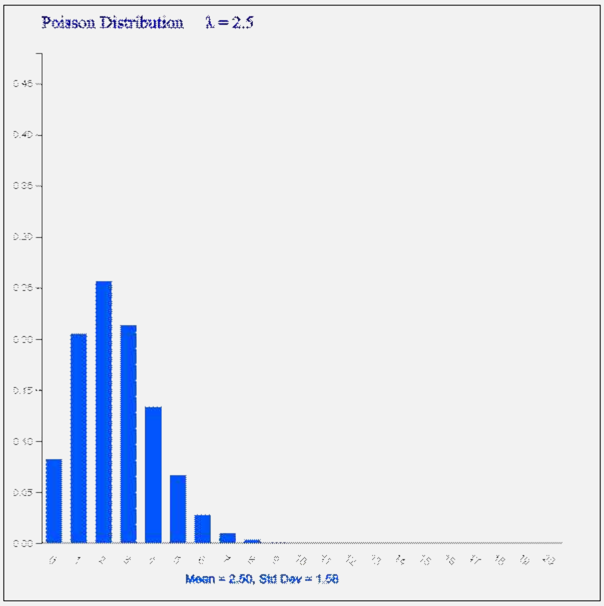
Example 5.3.9

Assume that the average number of Typhoons passing through the southern part of the country per year is a Poisson distribution with λ = 2.5. Check the following probabilities using 『eStatU』.

1) What is the probability that a Typhoon will pass once this year?
2) What is the probability that Typhoons will pass twice or three times or four times this year?
3) What is the probability that Typhoons will pass more than once this year?

Answer

- Select 'Poisson distribution' from the menu of 『eStatU』 and select λ = 2.5. Then click on the [Execute] button to display a graph such as <Figure 5.3.12> and click the [Poisson Prob Table] button to see the Table 5.3.7.



<Figure 5.3.12> 『eStatU』 Poisson Distribution when λ = 2.5

Table 5.3.7 Poisson Distribution when λ = 2.5

λ = 2.5			
x	P(X = x)	P(X $\le$ x)	P(X $\ge$ x)
0	0.0821	0.0821	1.0000
1	0.2052	0.2873	0.9179
2	0.2565	0.5438	0.7127
3	0.2138	0.7576	0.4562
4	0.1336	0.8912	0.2424
⋯	⋯	⋯	⋯

1) P(X=3) = 0.0821.
2) P(2 $\le$ X $\le$ 4) can be calculated as follows:

$$P(2 \le X \le 4) = P(X \le 4) - P(X \le 1) = 0.8912 - 0.2873 = 0.6039$$

This event can be calculated as P(X=2) + P(X=3) + P(X=4).

3) If you use Table 5.3.7, P(X $\ge$ 2) = 0.7127. Then the probability can be calculated by using the complementary event as follows:

$$P(X \ge 2) = 1 - P(X \le 1) = 1 - 0.2873 = 0.7127$$

[Practice 5.3.3]

The number of defects per 1 square meter of the fabric follows a Poisson distribution with the average number of defects λ = 0.2. When 1 square meter of the fabric is investigated for quality inspection, find the following probabilities using 『eStatU』.

1) What is the probability that the number of defects is zero?
2) What is the probability that the number of defects is greater than 2?

- Binomial distribution and Poisson distribution are very closely related. Mathematically, if n is very large and p is very small, the binomial distribution function converges to the Poisson distribution function. If you are interested in proofing this theory in detail, please refer to a book of mathematical statistics.

5.3.3 Geometric Distribution

- The binomial distribution is a probability distribution which counts the number of heads when a coin is thrown n times. On the other hand, the number of trials until the head of a coin appears may be of interest. Let us consider the following examples.

  - A candidate has 40 percent approval rating in an election. When interviewing voters to hear opinions, what is the probability of meeting one person who abstains from voting the candidate at the fifth trial?
  - The defect rate in products of a factory is said to be about 5%. If you continue to inspect the product until you find a defect product to investigate the cause, what is the probability of finding it at the 10^{th} trial?

- In these examples, as with the binomial distribution, we don't know what the outcome of each trial will be, but there are only two possible outcomes such as {pro, con} and {defective, normal}. If we denote one outcome of interest as ‘success’ and the other as ‘failure’, the experiment is the repetition of Bernoulli trials until we have one ‘success’. 'The number of Bernoulli trials until success' is called a **geometric random variable** and its distribution is called a **geometrical distribution**.
- The probability of success p in the geometrical distribution is called a parameter of the geometric distribution. The probability distribution function of the geometric distribution is as follows:

Definition

Geometric Distribution

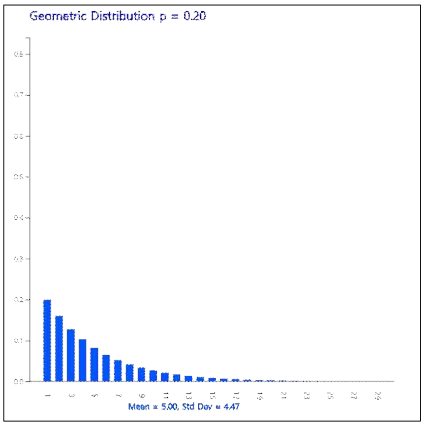
When the probability of 'success' in a Bernoulli trial is p and X is the number of Bernoulli trials until the first success, the probability distribution of X is called a **geometric distribution** and its probability distribution function is as follows:

$$f(x) = (1-p)^{x-1}p, \qquad x = 1,2, \ldots$$

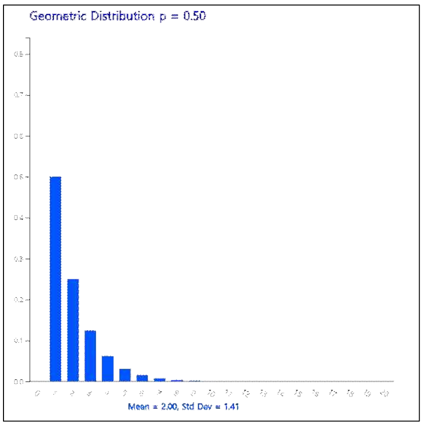
The expectation and variance of the geometric random variable are as follows:

$$E(X) = \frac{1}{p}, \quad V(X) = \frac{1-p}{p^2}$$

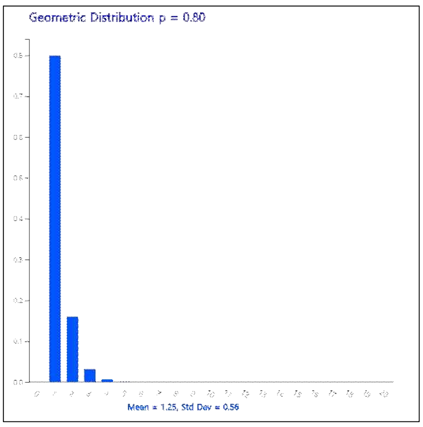
- <Figure 5.3.13> to <Figure 5.3.15> show the distribution of geometric distribution for different values of p.



<Figure 5.3.13> Geometric distribution when p = 0.2



<Figure 5.3.14> Geometric distribution when p = 0.5



<Figure 5.3.15> Geometric distribution when p = 0.8

Example 5.3.10

A candidate has 40% approval rating in an election. When interviewing voters to hear the opinions of those who oppose the candidate, calculate the following probabilities.

1) What is the probability of finding someone who is opposed in the first interview?
2) What is the probability of finding someone who is opposed in the fifth interview?

Answer

◆ Let X be the geometric random variable with p = 0.4.

1) P(X = 1) = f(1) = $(1-0.4)^{1-1}0.4$ = 0.4
2) P(X = 5) = f(5) = $(1-0.4)^{5-1}0.4$ = 0.0518

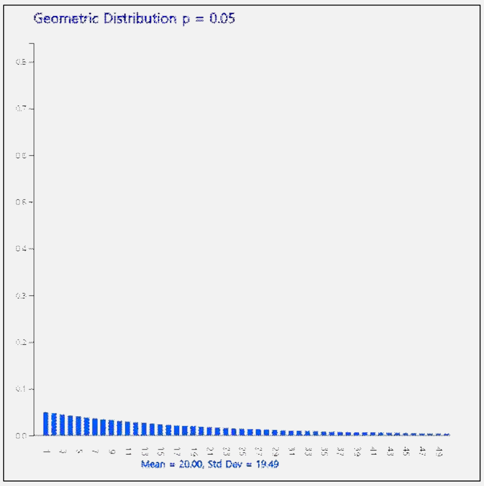
Example 5.3.11

The defect rate of a product produced by a factory is said about 5 percent. Use 『eStatU』 to obtain the following probabilities when continuing to inspect the product until it finds a defective product to investigate the cause of defective.

1) The probability of finding a defective product at the third trial.
2) The probability of finding a defective product at the third and more trial.

Answer

◆ Select 'Geometric Distribution' from the 『eStatU』 menu, select parameter p = 0.05, and click the [Execute] button to display the graph shown in <Figure 5.3.16>, and click the [Geometric Prob Table] button to display Table 5.3.8.



<Figure 5.3.16> Geometric distribution when $p=0.05$

Table 5.3.8 Part of Geometric distribution when $p=0.05$

p = 0.05			
x	P(X = x)	P(X $\leq$ x)	P(X $\geq$ x)
1	0.0500	0.0500	1.0000
2	0.0475	0.0975	0.9500
3	0.0451	0.1426	0.9025
4	0.0429	0.1855	0.8574
5	0.0407	0.2262	0.8145
…	…	…	…

Example 5.3.11 Answer (continued)

1) We can easily find P(X=3) = 0.0451.
2) We can easily find P(X $\geq$ 3) = 0.9025. We can use the complementary probability as follows:

P(X $\geq$ 3) = 1 - P(X $\leq$ 2) = 1 - 0.0975 = 0.9025

◆

[Practice 5.3.4]

The defect rate of products produced in a factory is about 1 percent. Use 『eStatU』 to obtain the following probabilities when continuing to inspect the product until it finds a defective product to investigate the cause of defective.

1) The probability of finding a defective product at the second trial.
2) The probability of finding a defective product at the third or more trial.

5.3.4 Hypergeometric Distribution

- Consider a statistical experiment to examine products in a factory to determine whether a box of the products includes defects or not. For example, consider a box consisting of 20 products and 15 of them are normal products and 5 are defective products. When three of the 20 products are sampled, the probability of having two normal products and one defective product can be calculated using the combination studied in section 5.1 as follows:

$$\frac{{}_{15}C_2 \times {}_{5}C_1}{{}_{20}C_3}$$

- A random variable that counts the number of 'success' in a finite population consisting of only 'success' and 'failure' is called a **hypergeometric random variable** and its distribution is called a **hypergeometric distribution**. The probability distribution function of the hypergeometric distribution is as follows:

Definition

Hypergeometric Distribution

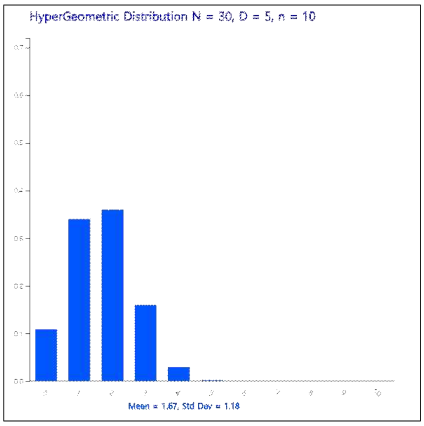
Consider a population of size N which consists of D 'success' and $N-D$ 'failure'. If we collect a sample of size n without replacement and X is the number of 'success' in the sample, then the distribution of X is called hypergeometric distribution and its probability distribution function is as follows:

$$f(x) = \frac{{}_{D}C_{x} \ {}_{N-D}C_{n-x}}{{}_{N}C_{n}}$$

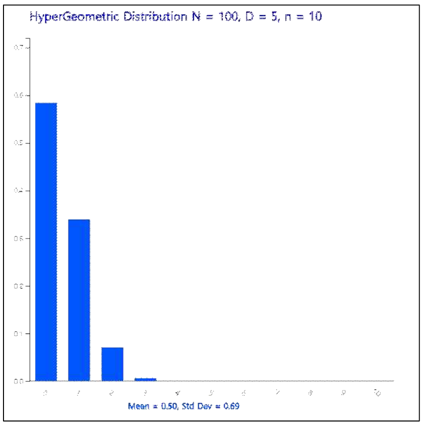
If we let $p = D/N$, the expectation and variance of the hypergeometric random variable are as follows:

$$E(X) = np, \ V(X) = np(1-p)\frac{N-n}{N-1} \cdots$$

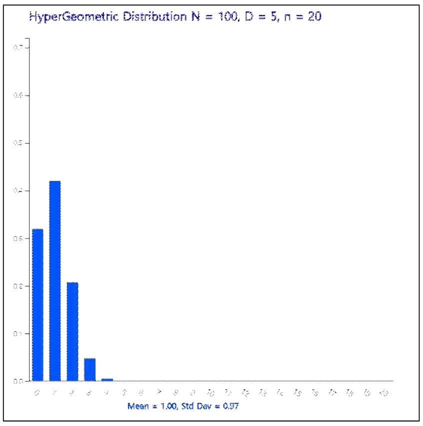
- In the hypergeometric probability distribution function, N, D, n are called parameters of the distribution and <Figure 5.3.17> to <Figure 5.3.19> show the distributions for various parameters.



<Figure 5.3.17> Hypergeometric distribution when N = 30, D = 5, n = 10



<Figure 5.3.18> Hypergeometric distribution when N = 100, D = 5, n = 10



<Figure 5.3.19> Hypergeometric distribution when N = 100, D = 5, n = 20

- Be aware that when the size of population N is very large or when we select the sample with replacement, the number of 'success' in the sample will follow the binomial distribution. If the size of population is finite and the sample is selected without replacement, the number of 'success' in the sample follows the hypergeometric distribution. Note that, when the number of products is finite and a selected product is not replaced, the failure rate changes.

Example 5.3.12

Sample of size 3 is selected from a box containing 20 tobacco products of which there are 15 normal products and 5 defective products. What is the probability of having one, two, or three defective products in the sample?

Answer

- These probability calculations have already been studied using combinations in section 5.1. This is the hypergeometric distribution with N = 20, D = 15, n = 3, so the probabilities are as follows:

$$P(X=1) = \frac{{}_{15}C_2 \times {}_5C_1}{{}_{20}C_3} = \frac{15 \times 10}{1140} = 0.460$$

$$P(X=2) = \frac{{}_{15}C_1 \times {}_5C_2}{{}_{20}C_3} = \frac{105 \times 5}{1140} = 0.132$$

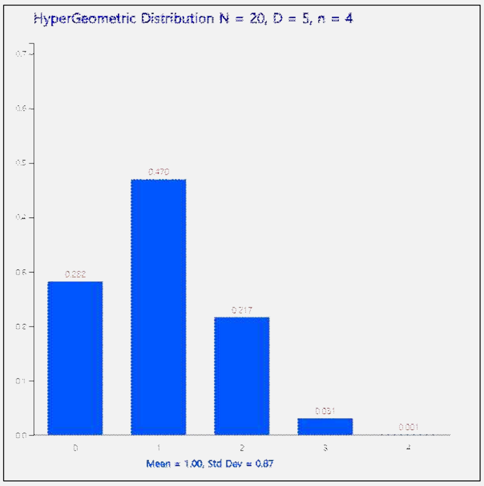
$$P(X=3) = \frac{{}_{15}C_0 \times {}_5C_3}{{}_{20}C_3} = \frac{455 \times 1}{1140} = 0.099$$

Example 5.3.13

Use 『eStatU』 to obtain the probability of [Example 5.3.12].

Answer

- Select 'Hypergeometric Distribution' from the menu of 『eStatU』, select N = 20, D = 15, n = 3 and click on the [Execute] button to display a graph such as <Figure 5.3.20>. If you click the [Hypergeometric Prob Table] button, Table 5.3.9 appears. This table shows the probabilities of P(X=0), P(X=1), P(X=2), and P(X=3).



<Figure 5.3.20 Hypergeometric distribution when N = 20, D=5, n=3

Table 5.3.9 Hypergeometric distribution when N = 20, D=5, n=3

N = 20	D = 5	n = 3	
x	P(X = x)	P(X ≤ x)	P(X ≥ x)
0	0.3991	0.3991	1.0000
1	0.4605	0.8596	0.6009
2	0.1316	0.9912	0.1404
3	0.0088	1.0000	0.0088

[Practice 5.3.5]

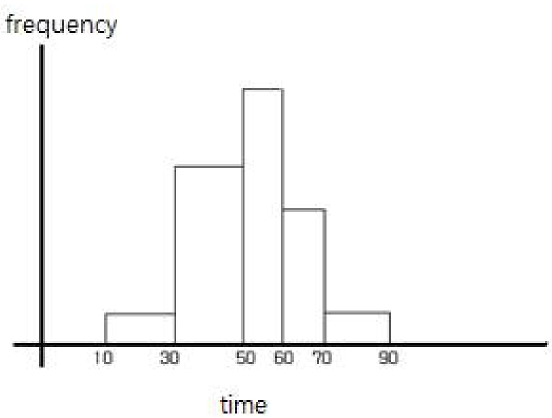
Sample of size 5 is selected from a box containing 20 cookie products of which there are 17 normal products and 3 defective products. Find the probability of having one, two, or three defective products in the sample using『eStatU』.

5.4 Continuous Random Variable

- Consider a statistical experiment that measures how long it takes for an office worker to get to work from home. Past experience shows that the commuting time usually takes about 30 minutes to get to the work place if the traffic is not congested. While the result of this experiment will have a real number near 30 minutes, we can assume generally that the sample space is larger than zero, and define a random variable X as the 'commuting time to work place'. As such, if a random variable has an infinite number of possible values and it is uncountable, it is called a **continuous random variable**.
- In case of the continuous random variable, calculating probability at each value of the random variable is meaningless, because there are infinite possible values and the probability at each value is considered zero. Instead of calculating the probability at a single value, the probability of an interval is of interest in case of the continuous random variable. For example, 'What is the probability of a commuting time between 25 and 35 minutes?' In order to obtain this probability, we can divide the sample space of the commuting time into several intervals and count the number of their frequencies and probabilities for 100 days as in Table 5.4.1. <Figure 5.4.1> is a histogram of this table.

Table 5.4.1 Frequency table of the commuting time for 100 days
X = ‘commuting time’

Interval ($a \le X < b$) unit: minute	Frequency	Probability
$10 \le X < 30$	5	5/100
$30 \le X < 50$	30	30/100
$50 \le X < 60$	40	40/100
$60 \le X < 70$	20	20/100
$70 \le X < 90$	5	5/100
Total	100 (days)	1



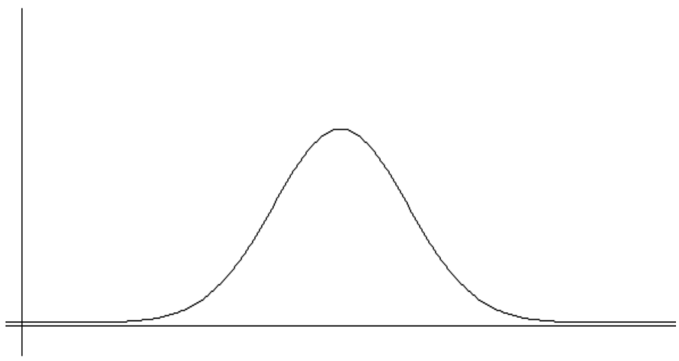
<Figure 5.4.1> Histogram of X = ‘commuting time’

- Using this frequency table, the probability of commuting time between 30 and 60 minutes can be calculated as follows:

  P(30 ≤ X < 60) = 30/100 + 40/100 = 70/100

- However, if you use this table, you can not calculate, for example, the probability of the commuting time between 25 and 35 minutes. In order to calculate this probability, calculation will require a detail frequency table and a histogram such as <Figure 5.4.2> which has narrower intervals by obtaining more data. If you increase the number of data and make the width of the interval close to zero, this histogram will be approximated to a continuous function as shown in <Figure 5.4.3>. This function is called a **probability distribution function of the continuous random variable**. As shown in this Figure, many real world data have a bell shape, large amount of data are observed near the mean, symmetrical about the mean. It is called a normal distribution.

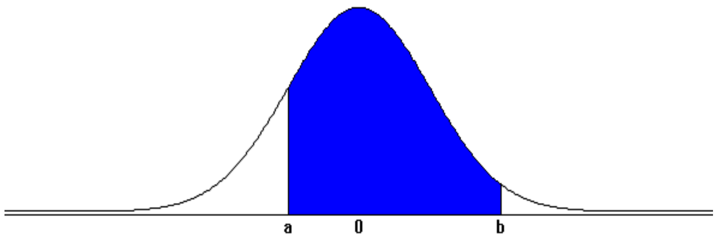
<Figure 5.4.2> Histogram with narrower intervals on many data



<Figure 5.4.3> Probability distribution function of a continuous random variable

- If the probability distribution function of a continuous random variable can be expressed as a mathematical function f(x), the desired probability can be obtained without finding the frequency table and histogram. The probability of the random variable X at interval (a, b), denoted as P(a < X < b), can be obtained as the area between (a, b) of f(x) as <Figure 5.4.4> which is the integral over (a, b) as follows:

$$P(a < X < b) = \int_a^b f(x)dx$$



<Figure 5.4.4> P(a < X < b) of the continuous random variable X

- The area under this function f(x) should be 1, because the addition of all probabilities is 1.

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

- The integral over $(a,\ b)$ of a function f(x) is generally difficult to obtain. For a normal distribution function, we use a table to calculate the probability which is discussed in Section 5.4.1. The following is an example to calculate the probability of the uniform distribution.

Example 5.4.1

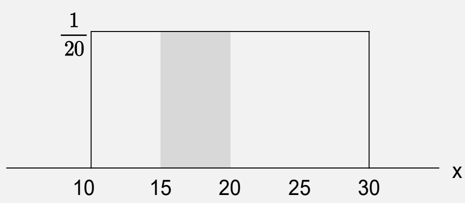
The delivery time to order a pizza and getting home have the same possibility as any time between 10 to 30 minutes (it is called a **uniform distribution)**. Let the random variable X be the time takes to deliver a pizza to home. Find a probability distribution function of X and draw a graph. Find the probability of the delivery time between 15 and 20 minutes.

Answer

- Since the random variable X has the same possibility as any number between 10 and 30, the probability distribution function (uniform distribution) is as follows:

$$f(x) = \begin{cases} 1/(30-10), & 10 < x < 30 \\ 0, & elsewhere \end{cases}$$

- <Figure 5.4.5> is the shape of this probability distribution function and it is called a uniform distribution between 10 and 30 denoted as Uniform(10,30).



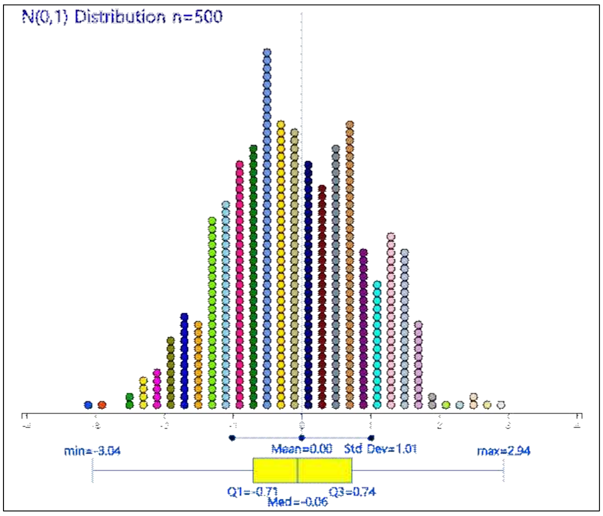
<Figure 5.4.5> Uniform distribution on (10,30) and the probability of P(15 < X < 20)

- The probability of the delivery time between 15 and 20 minutes is the area of the shaded rectangle of the <Figure 5.4.5> which can be calculated as follows:

P(15 < X < 20) = (20 - 15) × (1 / 20) = 0.25

5.4.1 Normal Distribution

- In real life, there are many continuous data that appears in the form of bell-shape as in <Figure 5.4.3>. The graph shows that large amount of data are located around their mean, fewer data located as it moves away from the mean, and is symmetrical around the mean. This type of data is called a normal distribution. Data obtained from measurements such as the height, weight, and length of bolt often follow the normal distribution. <Figure 5.4.6> shows a simulation of data which follow a normal distribution with the mean 0 and variance 1 using 『eStatU』.



<Figure 5.4.6> Simulation of data which follow a Normal distribution with mean 0 and variance 1 using 『eStatU』

- To make it easier to calculate probability for this type of data in the form of a normal distribution, many mathematicians tried to find a function to describe this distribution type. Abraham de Moivre (1667-1754) was the first who discovered the function, and then Carl Friedrich Gauss (1777-1855) extensively applied to physics and astronomy. This function is called a **normal distribution function** or a **Gaussian distribution function** and the functional form is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp[-\frac{(x-\mu)^2}{2\sigma^2}] \qquad -\infty < x < \infty$$

Definition

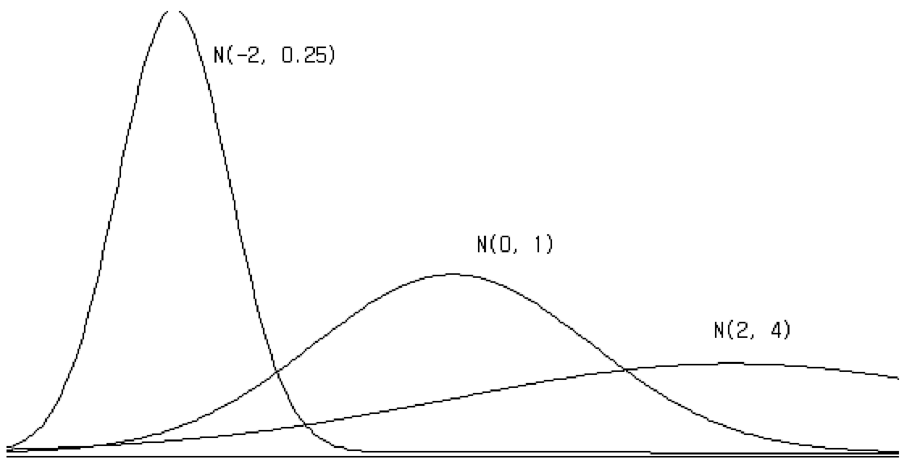
Normal Distribution

A normal distribution function or a Gaussian distribution function is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp[-\frac{(x-\mu)^2}{2\sigma^2}] \qquad -\infty < x < \infty$$

This distribution function has two parameters μ and σ, each representing the mean and standard deviation of the normal distribution.

- This distribution function has two parameters μ and σ, each representing the mean and standard deviation of the normal distribution. If X is a normal random variable with mean μ and variance σ^2, it is often denoted by a symbol X $\sim$ N (μ,σ^2). <Figure 5.4.7> shows three normal distributions, N(-2,0.25), N(0,1), and N(2,4) simultaneously.



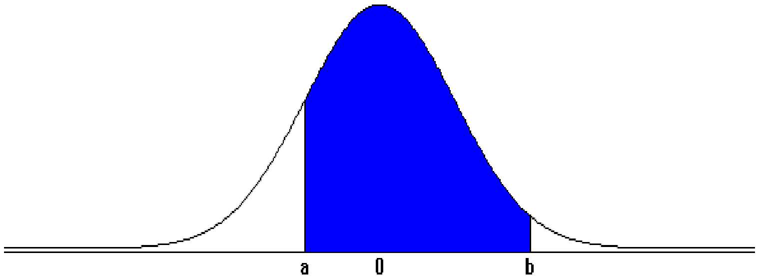
<Figure 5.4.7> Three Normal distribution N(-2,0.25), N(0,1), N(2,4)

- Characteristics of the normal distribution can be summarized as following.

1) It is a continuous function in the shape of a bell.
2) It is symmetrical with respect to the mean μ. So the probability of the left and right side of the mean is 0.5 each.
3) There are infinite number of normal distributions according to the value of μ and σ.
4) The probability of the interval $[\mu - \sigma, \mu + \sigma]$ is 0.68, and the probability of the interval $[\mu - 2\sigma, \mu + 2\sigma]$ is 0.95, and the probability of the interval $[\mu - 3\sigma, \mu + 3\sigma]$ is 0.997. It implies that the Normal random variable has the most of values (99.7%) around the interval of (mean) $\pm$ 3 (standard deviation) and there are a few values outside of this interval.

Probability Calculation of the Normal Distribution

- The normal distribution is the most frequently used distribution in statistics. If X is a normal random variable with mean μ and variance σ^2, it requires a probability calculation in the interval (a,b). As described earlier, the probability of X on the interval (a,b), $P(a < X < b)$, is the area of f(x) surrounding the X-axis and interval $(a,\ b)$ as shown in <Figure 5.4.8>.



<Figure 5.4.8> Probability of X on the interval $(a,\ b)$, $P(a < X < b)$

Mathematically, this area must be obtained with the following definite integral over $(a,\ b)$, but it is impossible to calculate by hand and can only be calculated using a computer.

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

- If X is a normal random variable with the mean μ and variance σ^2, a **standardized random variable** Z = (X - μ) / σ is a normal random variable with the mean 0 and variance 1, i.e., Z $\sim$ N(0,1). This fact implies that, if we can find probabilities of all types of intervals in N(0,1) distribution, then we can also find probabilities of all types of intervals in N(μ, σ^2). Therefore, N(0,1) is called a **standard normal distribution** or simply Z distribution.

Theorem 5.4.1

Distribution of the Standardized Normal Random Variable

If X is a normal random variable with the mean μ and variance σ^2, i.e. X $\sim$ N(μ,σ^2), then the standardized random variable Z,

$$Z = \frac{X-\mu}{\sigma}$$

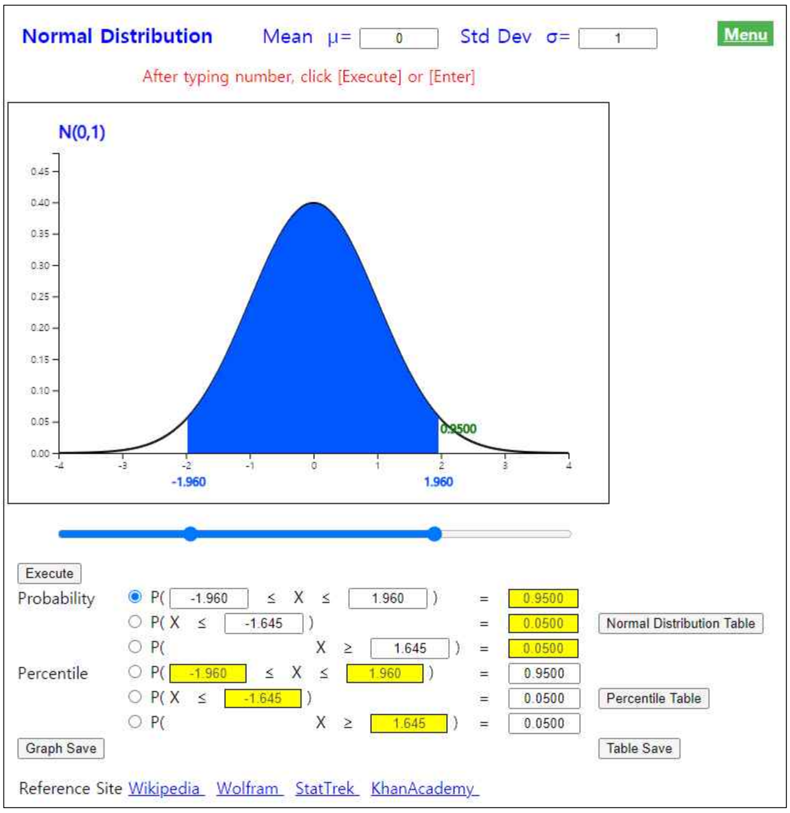
follows a Normal distribution with the mean 0 and variance 1, i.e.
Z $\sim$ N(0,1).

- For the standard normal distribution function, the probability P(Z < z) which is the area from the left end($-\infty$) to the value z is calculated by using a computer and summarized as Table 5.4.2. It is called the **standard normal distribution table** and Table 5.4.2 is a part of this table obtained using 『eStatU』. This table covers values of z between ⁻3.99 and 3.99 by increment of 0.01 and four decimal digits of the probability for an interval ($-\infty$, z) in the standard normal distribution is calculated. Calculation of the probability by using this table is usually enough approximation for practical application.

Table 5.4.2 Standard normal distribution table by using『eStatU』

Normal Distribution	μ = 0	σ = 1.000													
x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)
-3.99	0.0000	-2.99	0.0014	-1.99	0.0233	-0.99	0.1611	0.01	0.5040	1.01	0.8438	2.01	0.9778	3.01	0.9987
-3.98	0.0000	-2.98	0.0014	-1.98	0.0239	-0.98	0.1635	0.02	0.5080	1.02	0.8461	2.02	0.9783	3.02	0.9987
-3.97	0.0000	-2.97	0.0015	-1.97	0.0244	-0.97	0.1660	0.03	0.5120	1.03	0.8485	2.03	0.9788	3.03	0.9988
-3.96	0.0000	-2.96	0.0015	-1.96	0.0250	-0.96	0.1685	0.04	0.5160	1.04	0.8508	2.04	0.9793	3.04	0.9988
-3.95	0.0000	-2.95	0.0016	-1.95	0.0256	-0.95	0.1711	0.05	0.5199	1.05	0.8531	2.05	0.9798	3.05	0.9989
-3.94	0.0000	-2.94	0.0016	-1.94	0.0262	-0.94	0.1736	0.06	0.5239	1.06	0.8554	2.06	0.9803	3.06	0.9989
-3.93	0.0000	-2.93	0.0017	-1.93	0.0268	-0.93	0.1762	0.07	0.5279	1.07	0.8577	2.07	0.9808	3.07	0.9989
-3.92	0.0000	-2.92	0.0018	-1.92	0.0274	-0.92	0.1788	0.08	0.5319	1.08	0.8599	2.08	0.9812	3.08	0.9990
-3.91	0.0000	-2.91	0.0018	-1.91	0.0281	-0.91	0.1814	0.09	0.5359	1.09	0.8621	2.09	0.9817	3.09	0.9990
-3.90	0.0000	-2.90	0.0019	-1.90	0.0287	-0.90	0.1841	0.10	0.5398	1.10	0.8643	2.10	0.9821	3.10	0.9990
-3.89	0.0001	-2.89	0.0019	-1.89	0.0294	-0.89	0.1867	0.11	0.5438	1.11	0.8665	2.11	0.9826	3.11	0.9991
-3.88	0.0001	-2.88	0.0020	-1.88	0.0301	-0.88	0.1894	0.12	0.5478	1.12	0.8686	2.12	0.9830	3.12	0.9991
-3.87	0.0001	-2.87	0.0021	-1.87	0.0307	-0.87	0.1922	0.13	0.5517	1.13	0.8708	2.13	0.9834	3.13	0.9991
-3.86	0.0001	-2.86	0.0021	-1.86	0.0314	-0.86	0.1949	0.14	0.5557	1.14	0.8729	2.14	0.9838	3.14	0.9992
-3.85	0.0001	-2.85	0.0022	-1.85	0.0322	-0.85	0.1977	0.15	0.5596	1.15	0.8749	2.15	0.9842	3.15	0.9992
-3.84	0.0001	-2.84	0.0023	-1.84	0.0329	-0.84	0.2005	0.16	0.5636	1.16	0.8770	2.16	0.9846	3.16	0.9992
-3.83	0.0001	-2.83	0.0023	-1.83	0.0336	-0.83	0.2033	0.17	0.5675	1.17	0.8790	2.17	0.9850	3.17	0.9992
-3.82	0.0001	-2.82	0.0024	-1.82	0.0344	-0.82	0.2061	0.18	0.5714	1.18	0.8810	2.18	0.9854	3.18	0.9993
-3.81	0.0001	-2.81	0.0025	-1.81	0.0351	-0.81	0.2090	0.19	0.5753	1.19	0.8830	2.19	0.9857	3.19	0.9993
-3.80	0.0001	-2.80	0.0026	-1.80	0.0359	-0.80	0.2119	0.20	0.5793	1.20	0.8849	2.20	0.9861	3.20	0.9993
-3.79	0.0001	-2.79	0.0026	-1.79	0.0367	-0.79	0.2148	0.21	0.5832	1.21	0.8869	2.21	0.9864	3.21	0.9993
-3.78	0.0001	-2.78	0.0027	-1.78	0.0375	-0.78	0.2177	0.22	0.5871	1.22	0.8888	2.22	0.9868	3.22	0.9994
-3.77	0.0001	-2.77	0.0028	-1.77	0.0384	-0.77	0.2206	0.23	0.5910	1.23	0.8907	2.23	0.9871	3.23	0.9994
-3.76	0.0001	-2.76	0.0029	-1.76	0.0392	-0.76	0.2236	0.24	0.5948	1.24	0.8925	2.24	0.9875	3.24	0.9994

- In『eStatU』, the calculation of probability P(a < X < b) for the interval (a, b) of any normal distribution N(μ, σ^2) can be done as in <Figure 5.4.9>, and the percentile x for a given probability p, which is P(X < x) = p, can also be easily calculated. In『eStatU』, the probability of any interval on [$\mu - 4\sigma$, $\mu + 4\sigma$] can be calculated. The probability of P(Z < z) is near 0 if z is less than $\mu - 4\sigma$ and is 1 if z is greater than $\mu + 4\sigma$. Table 5.4.3 shows percentiles of the standard normal distribution by using『eStatU』.



<Figure 5.4.9> Normal probability calculation using『eStatU』

Table 5.4.3 Percentiles of standard normal distribution by using『eStatU』

Normal Distribution	μ = 0	σ = 1.000							
p	P(X ≤ x) = p	p	P(X ≤ x) = p	p	P(X ≤ x) = p	p	P(X ≤ x) = p	p	P(X ≤ x) = p
0.005	-2.576	0.205	-0.824	0.405	-0.240	0.605	0.266	0.805	0.860
0.010	-2.326	0.210	-0.806	0.410	-0.228	0.610	0.279	0.810	0.878
0.015	-2.170	0.215	-0.789	0.415	-0.215	0.615	0.292	0.815	0.896
0.020	-2.054	0.220	-0.772	0.420	-0.202	0.620	0.305	0.820	0.915
0.025	-1.960	0.225	-0.755	0.425	-0.189	0.625	0.319	0.825	0.935
0.030	-1.881	0.230	-0.739	0.430	-0.176	0.630	0.332	0.830	0.954
0.035	-1.812	0.235	-0.722	0.435	-0.164	0.635	0.345	0.835	0.974
0.040	-1.751	0.240	-0.706	0.440	-0.151	0.640	0.358	0.840	0.994
0.045	-1.695	0.245	-0.690	0.445	-0.138	0.645	0.372	0.845	1.015
0.050	-1.645	0.250	-0.674	0.450	-0.126	0.650	0.385	0.850	1.036
0.055	-1.598	0.255	-0.659	0.455	-0.113	0.655	0.399	0.855	1.058
0.060	-1.555	0.260	-0.643	0.460	-0.100	0.660	0.412	0.860	1.080
0.065	-1.514	0.265	-0.628	0.465	-0.088	0.665	0.426	0.865	1.103
0.070	-1.476	0.270	-0.613	0.470	-0.075	0.670	0.440	0.870	1.126
0.075	-1.440	0.275	-0.598	0.475	-0.063	0.675	0.454	0.875	1.150
0.080	-1.405	0.280	-0.583	0.480	-0.050	0.680	0.468	0.880	1.175
0.085	-1.372	0.285	-0.568	0.485	-0.038	0.685	0.482	0.885	1.200
0.090	-1.341	0.290	-0.553	0.490	-0.025	0.690	0.496	0.890	1.227
0.095	-1.311	0.295	-0.539	0.495	-0.013	0.695	0.510	0.895	1.254
0.100	-1.282	0.300	-0.524	0.500	0.000	0.700	0.524	0.900	1.282
0.105	-1.254	0.305	-0.510	0.505	0.013	0.705	0.539	0.905	1.311
0.110	-1.227	0.310	-0.496	0.510	0.025	0.710	0.553	0.910	1.341

Example 5.4.2

When Z is a standard normal random variable, find the following probability using standard normal distribution table. Then use『eStatU』to confirm the probability.

1) P(Z < 1.96)
2) P(-1.96 < Z < 1.96)
3) P(Z > 1.96)

Answer

◆ By using standard normal distribution table,

1) P(Z < 1.96) = 0.975.
2) P(-1.96 < Z < 1.96) = P(Z < 1.96) - P(Z < -1.96) = 0.975 - 0.025 = 0.95
3) P(Z > 1.96) = 1 - P(Z < 1.96) = 1 - 0.975 = 0.025

◆ By using normal distribution module of『eStatU』(<Figure 5.4.9>),

1) Enter 1.96 on the second option below the graph, then click the [Execute] button. The answer is shown at the yellow box in the right hand side

P(X ≤ 1.960) = 0.9750

2) Enter an interval from -1.96 and 1.96 on the first option below the graph, then click the [Execute] button.

P(-1.960 ≤ X ≤ 1.960) = 0.9500

3) Enter 1.96 on the third option below the graph, then click the [Execute] button. The answer is shown at the yellow box in the right hand side

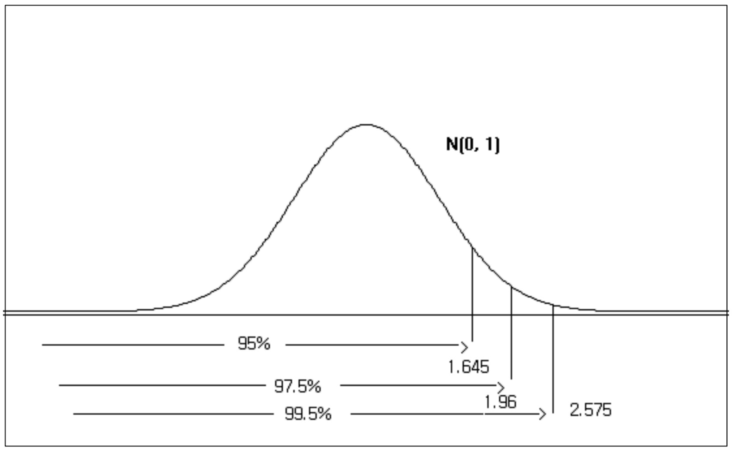
P(X ≥ 1.960) = 0.0250

Example 5.4.3	When Z is a standard normal random variable, obtain x that satisfies the following formula by using percentile table of the standard normal distribution. Then use 『eStat U』 to find this value x. 1) P(Z < x) = 0.90 2) P(-x < Z < x) = 0.99 3) P(Z > x) = 0.05
Answer	◆ By using percentile table of the standard normal distribution, 1) The value of x is 1.2826 2) The percentile of 0.995 is 2.575. 3) The value of x is 1.645. ◆ By using normal distribution module of 『eStatU』 as in <Figure 5.4.9>, 1) Enter p = 0.90 at the right box of the fifth option below the graph, then click the [Execute] button. It shows that the exact 90 percentile is 1.282 at the yellow box. ○ P(X ≤ 1.282) = 0.9000 2) Enter p = 0.99 at the right box of the fourth option below the graph, then click the [Execute] button. You can see that the exact two-sided value is −2.576 and 2.576 at the yellow boxes . ◉ P(-2.576 ≤ X ≤ 2.576) = 0.9900 3) Enter p = 0.05 at the right box of the sixth option below the graph, then click the [Execute] button. It shows that the right 5 percentile is 1.645 at the yellow box. ○ P(X ≥ 1.645) = 0.0500

[Practice 5.4.1]	When Z is a standard normal random variable, find following probabilities using 『eStat U』 . 1) Calculate the probability that Z is beween 0 and 1.5. 2) Calculate the probability that Z is between −1.5 and 0. 3) P(Z < -1.5) 4) P(Z > 1.5) 5) P(Z > -1.5) 6) P(Z < 1.5) 7) P(-1.5 < Z < 1.5)

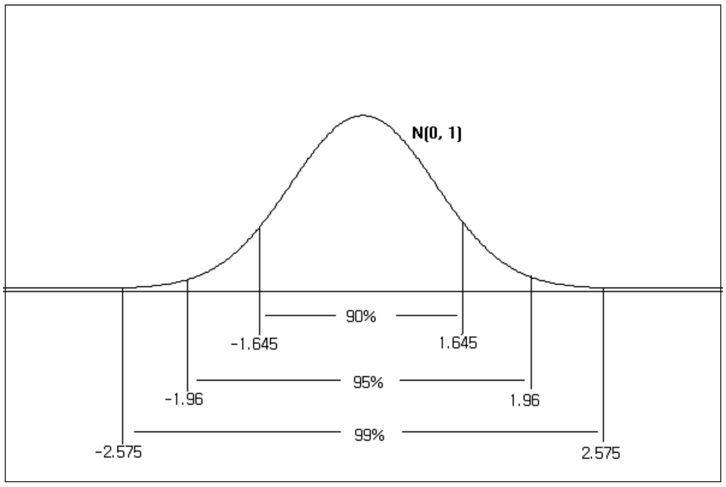
[Practice 5.4.2]	When Z is a standard normal random variable, find x that satisfies the following formula using 『eStatU』 . 1) P(Z < x) = 0.80 2) P(-x < Z < x) = 0.80 3) P(Z > x) = 0.80

- It is recommended that you remember probabilities of some intervals in the standard normal distribution which are frequently used. <Figure 5.4.10> shows percentiles that cumulated probabilities become 95%, 97.5%, and 99.5% from the left end of the standard normal distribution.



<Figure 5.4.10> Percentile with a cumulative probability of 95%, 97.5%, and 99.5% from the left end of N(0,1)

- <Figure 5.4.11> shows a value of 95% and 99% when two ends are excluded equally.



<Figure 5.4.11> Percentile with a probability of 95%, 97.5%, and 99.5% when the two ends are excluded equally in N(0,1).
i.e., P(-1.96<Z<1.96) = 0.95, P(-2.575<Z<2.575) = 0.99

- The probability of a normal distribution in general can be obtained by using standard normal distribution table. As we studied, if X is a normal random variable with the mean μ and variance σ^2, $Z = (X-\mu)/\sigma$ follows the standard normal distribution. Therefore, the probability P(a < X < b) of the interval (a,b) of X can be obtained from the standard normal distribution as follows:

Theorem 5.4.1

Calculation of Normal Probability

When X is a normal random variable with a mean μ and variance σ^2, $Z = (X-\mu)/\sigma$ follows the standard normal distribution. Therefore, the probability P(a < X < b) of the interval (a,b) of X is as follows:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

Example 5.4.4

If mid-term scores (X) of a statistics course follow a normal distribution with the average of 70 points and the standard deviation of 10, calculate following probabilities. Check the calculated values by using 『eStatU』.

1) P(X < 94.3) 2) P(X > 57.7) 3) P(57.7 < X < 94.3)

Answer

- By using transformation to the standardized normal random variable, probability calculations are as follows:

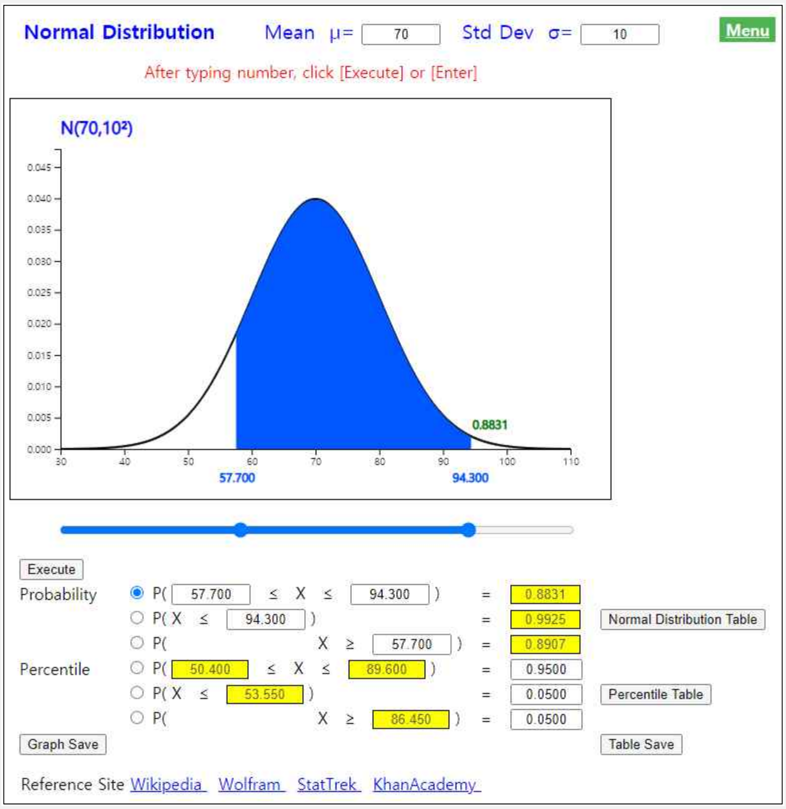
1) $\mathrm{P}(\mathrm{X} < 94.3) = \mathrm{P}(\frac{\mathrm{X}-70}{10} < \frac{94.3-70}{10}) = \mathrm{P}(\mathrm{Z} < 2.43) = 0.9925$

2) $\mathrm{P}(\mathrm{X} > 57.7) = \mathrm{P}(\frac{\mathrm{X}-70}{10} > \frac{57.7-70}{10}) = \mathrm{P}(\mathrm{Z} > -1.23) = 0.8907$

3) $\mathrm{P}(57.7 < \mathrm{X} < 94.3) = \mathrm{P}(\frac{57.7-70}{10} < \frac{\mathrm{X}-70}{10} < \frac{94.3-70}{10})$
$= \mathrm{P}(-1.23 < \mathrm{Z} < 2.43) = 0.8832$

- By using 『eStatU』, to obtain a probability of the normal distribution $N(70, 10^2)$, enter the mean as 70 and the standard deviation as 10 at the top of the screen as <Figure 5.4.12>.

1) Enter 94.3 at the second option below the graph and then click the [Execute] button.



<Figure 5.4.12> Probability calculation of $N(70, 10^2)$ distribution

2) Enter 57.7 at the third option below the graph and then click the [Execute] button.

3) Enter the interval as [57.7, 94.3] at the first option below the graph and then click the [Execute] button.

Example 5.4.5

In [Example 5.4.4], obtain the following percentiles by using normal probability table and also by using 『eStatU』.

1) What is the 95% percentile of the mid-term test scores?
2) What is the 95% percentile of two-sided type of the mid-term scores?

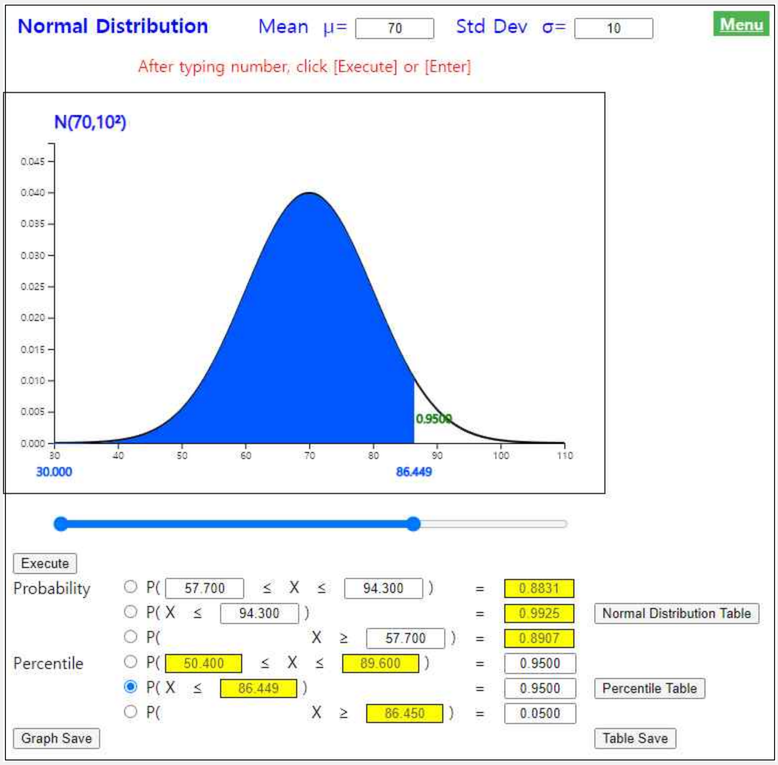
Answer

◆ By using normal probability table, percentile calculations are as follows:

1) The 95 percentile, $\mathrm{P}(Z < ?) = 0.95$, in the N(0,1) probability table is 1.645, so the percentile in the $N(70, 10^2)$ is 70 + 1.645 × 10 = 86.45.
2) The 95 percentile of two-sided type, $\mathrm{P}(? < Z < ?) = 0.95$, in the N(0,1) table, can be calculated from $\mathrm{P}(Z < ?) = 0.975$ which is 1.96. So the two-sided 95% percentile interval is [70 − 1.96 × 10, 70 + 1.96 × 10], i.e., [50.4, 89.6].

◆ By using 『eStatU』, to obtain the probability of normal distribution $N(70, 10^2)$, enter the mean as 70 and the standard deviation as 10 at the top of the screen as <Figure 5.4.13>.

1) Enter 0.95 at the box of the fifth option below the graph and click the [Execute] button to display the 95 percentile as 86.449.



<Figure 5.4.13> Percentile calculation of $N(70, 10^2)$ distribution

2) Enter 0.95 at the box of the fourth option below the graph and click the [Execute] button to display the two-sided 95 percentile as [50.400, 89.600].

[Practice 5.4.3]

The length of time a customer waits to receive a service at a bank follows a normal distribution with an average of 5 minutes and standard deviation of 1 minute. Calculate the following probability using『eStatU』.

1) The probability that a customer waits between 2 and 3 minutes.
2) The probability that a customer will wait less than 1 minute.
3) The probability that a customer waits at least 7 minutes to receive a service.

[Practice 5.4.4]

If total scores earned by students in a Statistics course follow a normal distribution with the average of 75 points and standard deviation of 10, find the following scores using『eStatU』.

1) Top 10% of the scores that will get A grade. What is the minimum score to get A grade?
2) Top 30% to 10% of the scores will get B grade. What is the score range to get B grade?

Normal Approximation of Binomial Distribution

- In case of large n (approximately 100 or more) in a binomial distribution with n and p, a direct probability calculation is not possible. In such cases, a normal distribution with the average np and variance np(1-p) is used to calculate an approximated probability as shown in the following example.

Example 5.4.6

The defect rate of products produced in a factory is 5 percent. One day, a sample of 100 products was collected. Answer the following questions.

1) What is the probability that there are less than two defective products?
2) What is the probability that there are defective products between 3 and 7?

Answer

- If the number of defective products is X, X is a binomial distribution with n = 100, p = 0.05. When n is this large, we calculate the probability approximately using normal distribution. Since the mean of this binomial distribution is np = 100 × 0.05 = 5, and the variance is np(1-p) = 100 × 0.05 x (1-0.05) = 4.75, we use the normal distribution N(5, 4.75) to calculate the probability approximately as follows:

1) $P(X \le 2) = P(Z \le \frac{(2-5)}{\sqrt{4.75}}) = P(Z \le -1.376) = 0.0845$

2) $P(3 \le X \le 7) = P(\frac{(3-5)}{\sqrt{4.75}} \le Z \le \frac{(7-5)}{\sqrt{4.75}})$

$= P(-0.918 \le Z \le 0.918) = 0.642$

5.4.2 Exponential Distribution

- Most of continuous data obtained in real life follows a normal distribution, but sometimes it is not. Let's take a look at the following examples.

  - Inter-arrival time of coming calls between 9 am and 10 am in an office.
  - Time interval between defective products appearing in a factory production line.

- These examples appear when events occur at the same rate at a given time (e.g., three calls per hour, etc.). If the average number of events per unit hour is λ and X is the random variable of the time between events, then X is an **exponential random variable**. λ is a parameter of the exponential distribution and the formula for the exponential probability distribution function is as follows:

Definition

Exponential Distribution

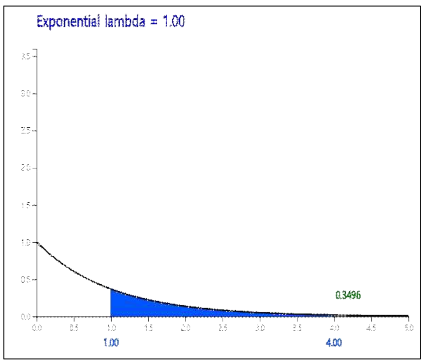
When the average number of events per unit hour is λ and the random variable X is the time between events, the probability distribution function of X is as follows:

$$f(x) = \lambda \exp(-\lambda x), \qquad x > 0$$

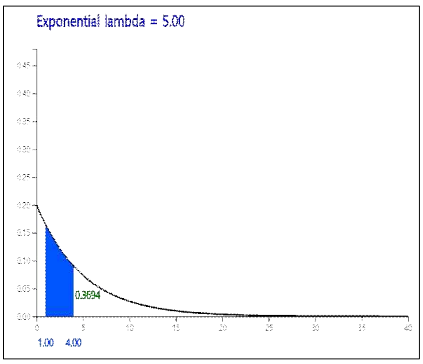
It is called an **exponential distribution** and its expectation and variance are as follows:

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}$$

- The exponential distribution is similar to the geometric distribution in discrete probability distributions. <Figure 5.4.14> and <Figure 5.4.15> show the exponential distribution function for different parameters.



<Figure 5.4.14> Exponential distribution when λ = 1.0



<Figure 5.4.15> Exponential distribution when λ = 5.0

- 『eStatU』 can easily calculate a probability for various values of the exponential distribution.

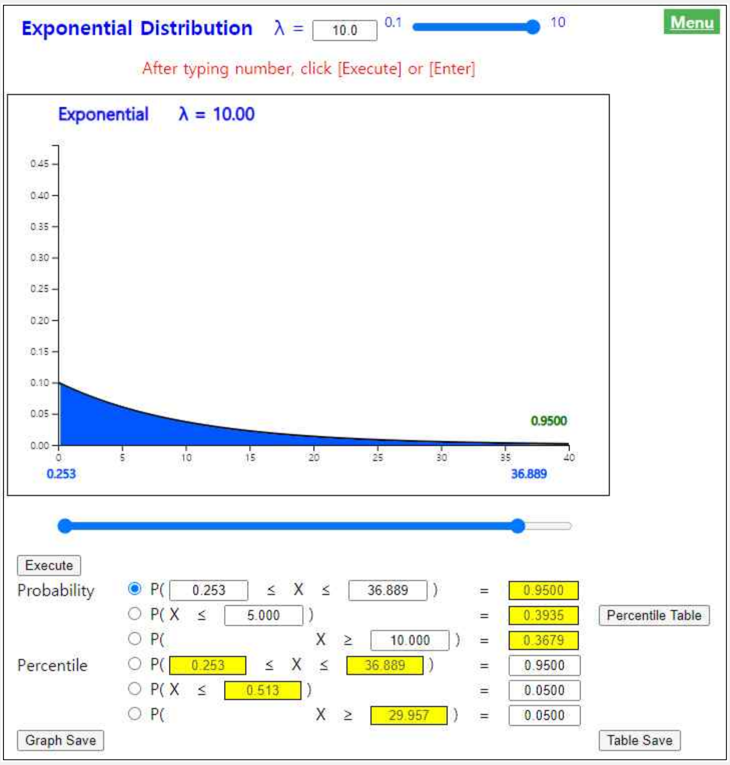
Example 5.4.7

If the life span of a product has the average of 10 hours and follows an exponential distribution, obtain the following probabilities using 『eStatU』.

1) What is the probability of a product having a lifespan of less than 5 hours?
2) What is the probability of a product having a lifespan more than 10 hours?

Answer

- In 『eStatU』, select 'Exponential Distribution' and enter λ = 10. Click the [Execute] button to reveal the graph shown as <Figure 5.4.16>.



<Figure 5.4.16> Exponential distribution when λ=10.0

1) Enter 5 at the box of the second option below the distribution graph as <Figure 5.4.16> and click the [Execute] button. The probability is 0.3935.

2) Similarly, enter 10 at the box of the third option below the distribution graph as <Figure 5.4.16> and click the [Execute] button. The probability is 0.3679.

[Practice 5.4.5]

A product has an average life expectancy of 1000 hours and follows an exponential distribution. Use 『eStatU』 to obtain the following probabilities.

1) What is the probability that the lifespan of the product is less than 700 hours?
2) What is the probability that the product has a lifespan of 1000 hours or more?

Exercise

5.1 Calculate the following.

1) ${}_8P_2$ 2) ${}_5P_3$ 3) ${}_9P_9$ 4) ${}_{10}P_4$ 5) ${}_6P_2$
6) ${}_{10}C_3$ 7) ${}_{10}C_7$ 8) ${}_7C_3$ 9) ${}_5C_1$ 10) ${}_8C_4$

5.2 When four offices are located side by side, they are run by four middle-level executives. How many ways are these four executives assigned to four offices?

5.3 A manager employs seven employees and four of them are to form a production team. How many ways can seven people form different production teams?

5.4 An advertising designer wants to design pages of a magazine by selecting three of eight photos. Page location is said to be unimportant. How many designs can be made using different combinations of photos?

5.5 The president of a company that produces five kinds of soap wants to display each sample soap in a row at a show case in his office. How many ways can you display five soaps?

5.6 A salesperson has seven items and wants them on display at the counter. He can only display four. If the order in which he displays the item is not important, how many ways can he display?

5.7 An airline company which has six airplanes plans to advertise each airplane's operation in the Sunday newspaper for six weeks.

1) How many ways are there to advertise flight operations?
2) If you have decided to advertise only four times, how many ways are there?

5.8 A company said that five applicants have applied to fill five different positions. If all applicants are given equal qualifications for five positions, how many ways are there to fill five positions?

5.9 A company which has 231 employees and employees are classified by age and rank as the following table.

	Age					
Rank	A1 < 20	A2 21-25	A3 26-30	A4 31-35	A5 > 35	Total
B1 Clerk	**20**	**20**	**15**	**10**	**5**	**70**
B2 Manger	**3**	**6**	**3**	**2**	**1**	**15**
B3 Technician	**15**	**30**	**35**	**20**	**10**	**110**
B4 Salesman	**1**	**5**	**10**	**5**	**2**	**23**
B5 Director	**0**	**1**	**5**	**2**	**0**	**8**
B6 Executive	**0**	**0**	**2**	**2**	**1**	**5**
Total	**39**	**62**	**70**	**41**	**19**	**231**

Refer to this table to explain the meaning of the following sets and find the number of each employee.

1) B1 ∩ A5 2) A2 ∩ B6 3) B4 ∩ A5 4) A1 ∪ B6 5) A3 ∪ A5
6) B2 ∪ B3 7) A4 8) (A1 ∪ A2) ∩ B3 9) (B3 ∪ B4) ∩ A5

What is the number of employees who meet each of the following conditions?

10) The number of persons who are neither Director nor Executive.
11) The number of persons in charge of both Director and Executive.
12) The number of persons aged older than 30 and in charge of Clerk or Manager.
13) The number of persons who are both Salesman and 21-25 years of age. The number of persons who are either Salesman or 21-25 years of age.
14) The number of Technicians under the age of 35.
15) The number of persons between the ages of 21 and 30 who are Technician or Salesman.
16) The number of persons who are Clerk or Manager and are over 30.

5.10 A store which has three types of goods, A, B and C, has received orders from 200 customers for a certain period of time.

Goods A: 100 Goods A and C: 55
Goods B: 95 Goods B and C: 30
Goods C: 85 Goods A and B and C: 20
Goods A and B: 50 people

Find the number of customers ordered using the Van diagram as follows:

1) Number of customers who have ordered at least one product.
2) Number of customers who have not ordered any product.
3) The exact number of customers who ordered one product.
4) The exact number of customers who have ordered two products.
5) The exact number of customers who have ordered three products.

5.11 Two executive positions in a company are vacant. Seven men and three women are fully qualified and equally eligible. The company decided to pick two of these ten people at random. Find the following odds.

1) The probability that a woman will be selected for both seats.
2) The probability that at least one of the two seats will be selected by a woman.
3) The probability that neither of the two seats will be selected by a woman.

5.12 There are 300 households living in a neighborhood. One hundred of them are said to be out of the house in the evening. Of the remaining households, 50 do not respond to telephone surveys. When a person randomly selects a household in the neighborhood and conducts a telephone survey, ask for the following odds.

1) Probability that no one will be present when an investigator calls a house.
2) When an investigator calls a house, there is someone, but there is a chance that they will not respond to the investigation.
3) Probability of responding to a telephone survey.

5.13 When an employee is selected randomly by a company, there is a 0.58 chance that the person is over the age of 31. What is the probability that an employee who is hired randomly is under 30?

5.14 The following table shows the number of absent days due to diseases of 100 workers over the year in a company. Answer the following questions based on this table.

X (Number of absent days)	0	1	2	3	4	5	6	7	8	9	10
Number of workers	5	8	10	12	18	14	10	9	8	4	2

1) Display the probability distribution of absent days due to disease.
2) Draw the cumulative probability distribution.
3) Calculate the expectation and variance.

5.15 When selecting one worker at random in question 5.14, obtain the following probabilities.

1) Probability that the selected person was absent for 3 days.
2) Probability that the selected person has been absent for more than five days.
3) Probability that the selected person was absent from 6 to 8 days.

5.16 In question 5.14, obtain the following.

1) $P(X = 0)$ 2) $P(X = 10)$ 3) $P(X \geq 6)$ 4) $P(X < 6)$
5) $P(3 \leq X \leq 7)$

5.17 A department store surveyed 80 customers about their purchasing habits and asked, 'How many times did you buy at this department store last month?' The answer to the question is as follows:

X(Number of purchases)	0	1	2	3	4	5	6	7
Number of customers	15	27	14	12	6	4	1	1

1) If X = 'Number of purchases', draw the probability distribution function of X.
2) Draw the cumulative distribution function of X.
3) When selecting one of customers, obtain the following probability.
   ① Probability that a person who purchased more than once will be selected.
   ② Probability that a person who has never purchased will be selected.
   ③ Probability that a person who has purchased more than 4 times will be selected.
   ④ Probability that a person who purchased less than 3 will be selected.
4) Find the expectation and variance of X.

5.18 Over the years, a salesperson found that there was a 50% chance of selling the product when a customer visited. If one day five customers visit this salesman, calculate the following odds. Check the calculation using 『eStatU』.

1) Exactly how likely is it to sell three product?
2) What is the probability of selling three or more products?
3) What is the probability of selling less than 3 products?
4) What is the probability that none of products could be sold?
5) What is the probability of selling 5 products?

5.19 If a random variable X follows a binomial distribution with $n = 6$, $p = 0.2$, find the following probabilities using 『eStatU』.

1) $P(X > 2)$ 2) $P(X \leq 4)$ 3) $P(2 \leq X \leq 4)$

5.20 In some cities, 35% of residents are said to be opposed to the idea of expanding the main street. Obtain the following probabilities and check them using 『eStatU』.

1) Probability that the number of residents who oppose the idea of expansion is 10 or more.
2) Probability that the number of residents who oppose the idea of expansion is between 15 and 18.
3) Probability that the number of residents who oppose the idea of expansion is less than 8.
4) Probability that the number of residents who oppose the idea of expansion is at least 12.
5) The probability that the number of residents who oppose the idea of expansion is only 13.

5.21 It is said that 72% of drivers usually use a seat belt. When you select 15 drivers randomly and calculate the following probabilities of drivers who use a seat belt and check the calculation using 『eStatU』.

1) Probability of 10 or more drivers.
2) 8 or less probability drivers.
3) Probability of at least 11 drivers.
4) Probability of at least 7 drivers.

5.22 One producer claims that 6% of his product is defective. If you sample 20 products randomly, what is the probability that the number of defective products is as follows. Check the calculation using 『eStatU』.

1) 2 defective products   2) more than 2($\geq$) products   3) 0 product
4) less than 5 products   5) between 2 and 5 products

5.23 The defect rate of products in a factory is said to be about 3%. When an employee continues to inspect the product until he finds a defective product to investigate the cause of defective products, find the following probabilities and check them using 『eStatU』.

1) The probability of finding a defective product at the third trial.
2) The probability of finding a defective product at least the third trial?

5.24 The number of defects per $1m^2$ in a fabric follows a Poisson distribution with the average defect λ = 0.2. When $1m^2$ of this fabric is investigated for quality inspection, find the following probabilities and check the calculation using 『eStatU』.

1) What is the probability that the defect count is zero?
2) What is the probability that the defect count is greater than 2?

5.25 Assume that the average number of Hurricanes passing through the southern part of the country is λ = 3 times per year. Check the following probabilities using 『eStatU』.

1) What is the probability that a Hurricane will pass once this year?
2) What is the probability that more than 2 Hurricanes will pass this year?
3) What is the probability that two or three Hurricanes will pass this year?

5.26 A candidate has a 60% approval rating in an election. When interviewing voters to hear the opinions of those who oppose the candidate, look for the next probabilities.

1) What is the probability of finding someone who is opposed in the first interview?
2) What is the probability of finding someone who is opposed in the fifth interview?

5.27 Five samples are selected randomly from a box containing 20 tobacco products of which there are

15 normal products and 5 defective products. What is the probability of having one, two, or three defectives in the samples?

5.28 If Z is the standard normal random variable, obtain the following and check the calculation using 『eStatU』.

1) Calculate an area between Z = 0 and Z = 1.54.
2) Calculate the probability that Z is between -2.07 and 2.33.
3) $P(Z \geq 0.65)$
4) $P(Z \geq -0.65)$
5) $P(Z < -2.33)$
6) $P(Z < 2.33)$
7) $P(-1.96 \leq Z \leq 1.96)$
8) $P(-2.58 \leq Z \leq 2.58)$
9) $P(-3.10 \leq Z \leq 1.25)$
10) $P(1.47 \leq Z \leq 3.44)$

5.29 It is said that a tin can lid produced by a company follows a normal distribution with the average diameter of 4 and the standard deviation of 0.012. Calculate the probability that the lids are between 3.97 and 4.03. Check the calculation using 『eStatU』.

5.30 It is said that the weight of a melon follows a normal distribution with the average of 250g and standard deviation of 12g. Find the probability that the weight of the melon is less than 260 grams. Check the calculation using 『eStatU』.

5.31 A bank employee found that the length of waiting time of a customer to receive a service follows a normal distribution with the average of 3 minutes and the standard deviation of 1 minute. Calculate the following probabilities and check the calculation using 『eStatU』.

1) Probability that a customer will wait between 2 and 3 minutes.
2) Probability that a customer will wait less than 1 minute.
3) Probability that a customer will wait at least 5 minutes.

5.32 It takes an average of 10 minutes for a production worker to finish a job at a factory and time follows a normal distribution with a standard deviation of 3 minutes. Calculate the following probabilities and check the calculation using 『eStatU』.

1) Probability of workers completing the work in 4 minutes.
2) Probability of workers completing the work at least 5 minutes.
3) Probability of workers completing the work within 3 minutes.

5.33 An average product has a life expectancy of 100 hours and follows an exponential distribution. Calculate the following probabilities and check the calculation using 『eStatU』.

1) What is the probability that a product has a lifespan less than 50 hours?
2) What is the probability that a product has a lifespan of 120 hours or more?

Multiple Choice Exercise

5.1 When two events A and B are mutually exclusive, what is the probability of an event A∪B?

① P(A)－P(B) ② P(A)P(B)
③ P(A)＋P(B) ④ P(A)/P(B)

5.2 Let the probability that event A will occur be P(A) and the probability that event B will occur be P(B). Which of the following is wrong?

① 0≤P(A)≤1 ② -1≤P(B)≤0
③ P(A∪B)=P(A)+P(B)-P(A∩B) ④ P(A∩B)=P(A) · P(B)

5.3 The value of P(A)=0.4, P(B)=0.2, P(A|B)=0.6. What is the probability of P(A∩B)?

① 0.08 ② 0.12
③ 0.24 ④ 0.48

5.4 If A ⊂ B, what is the comparison between the conditional probability P(A|B) and P(A)?

① equal to or greater ② smaller
③ equal or smaller ④ There is no comparison.

5.5 How likely are 2 and 5 to appear at the same time when throwing 2 dice?

① $\frac{1}{3}$ ② $\frac{1}{6}$
③ $\frac{1}{12}$ ④ $\frac{1}{18}$

5.6 When throwing a dice three times, what is the probability that the number of eyes is 5 at the first throw, 3 at the second, and an even number of eyes at the third?

① $\frac{1}{30}$ ② $\frac{1}{72}$
③ $\frac{1}{108}$ ④ $\frac{1}{276}$

5.7 When you extract three light bulbs randomly without replacement one by one from a barrel containing five good bulbs and two defective bulbs, what is the probability that one bulb is a defective item?

① $\frac{1}{7}$ ② $\frac{2}{7}$
③ $\frac{3}{7}$ ④ $\frac{4}{7}$

5.8 When P(B)=0.2, $\mathrm{P}(\mathrm{A} \cap \mathrm{B})$= 0.12, what is the value of P(A|B)?

① 0.20 ② 0.60
③ 0.12 ④ 0.24

5.9 When P(A)=0.4, P(B)=0.2, P(A|B)=0.6, what is the value of P(B|A)?

① 0.08 ② 0.24
③ 0.30 ④ 0.40

5.10 Mark the result of throwing two dice as (x_1, x_2) and let $B=\{(x_1,x_2)\mid x_1 > x_2\}$. What is the value of P(B)?

① $\frac{1}{3}$ ② $\frac{1}{12}$
③ $\frac{5}{12}$ ④ $\frac{1}{36}$

5.11 Mark the result of throwing two dice as (x_1, x_2) and let $A=\{(x_1,x_2)$ | x_1+x_2=10}, B={(x_1,x_2) | x_1> x_2}. What is the value of P(B|A)?

① $\frac{1}{3}$ ② $\frac{1}{12}$
③ $\frac{5}{12}$ ④ $\frac{1}{2}$

5.12 What is the standard deviation when we multiply five times each of the values of one random variable?

① It is five times as much. ② It is one-fifth the size.
③ It is ten times as much. ④ Four-tenths the size.

5.13 In a class, the average of mid-term scores was 24 points and its standard deviation was 3. Because of poor grades, professor doubled the mid-term scores and then add 10 points for all the students in the class. What is the average and standard deviation of new scores?

① 24, 3 ② 48, 3
③ 48, 6 ④ 58, 6

5.14 If the mean of a random variable X is 20, what is the mean of Y=2X+3?

① 20 ② 40
③ 23 ④ 43

5.15 If the variance of a random variable X is 2, what is the variance of Y=2X+1?

① 2 ② 3
③ 8 ④ 18

5.16 Which of the following is wrong?

① E(aX+b) = aE(X)+b ② E(X+Y) = E(X)+E(Y)
③ Var(aX+b) = a^2Var(X) ④ Var(aX+b) = a^2Var(X)+b

5.17 If the expected value of X is E(X)=5 and E(X^2)=25, and what is the variance V(X)?

① -5 ② 0
③ 5 ④ 25

5.18 Which one of the followings is suitable as a probability distribution function?

① $f(x) = \frac{x}{5}$ $(x = -1, 0, 1, 2, 3)$　　② $f(x) = \frac{(5-x)}{14}$ $(x = 0, 1, 2, 3)$
③ $f(x) = x^2 - 1$ $(x = 0, 1, 2, 3)$　　④ $f(x) = x^3 - x^2 + 5$ $(x = 0, 1, 2, 3)$

5.19 Which of the followings is NOT a probability distribution function?

① $f(x) \geq 0$　　② $\int_{-\infty}^{\infty} f(x)dx$=1
③ $P(a < X < b)$ =$\int_a^b f(x)dx$　　④ $\int_{-\infty}^{\infty} xf(x)\,dx$=1

5.20 What is the average value of X when the variables X take values of 0, 1, 2 and 3 and the probability function f(x) =$\frac{x}{6}$?

① $\frac{5}{3}$　　② $\frac{6}{3}$
③ $\frac{7}{3}$　　④ $\frac{8}{3}$

5.21 Which of the following is incorrect for the probability distribution function f(x) defined for all values of a continuous random variable X?

① f(x)≥0　　② $\int_{-\infty}^{\infty}$ f(x)dx=1
③ P(a≤x≤b)=$\int_a^b$f(x)dx　　④ $\int_0^{\infty}$f(x)dx=$\frac{1}{2}$

5.22 What is the mean value of E(X) and the standard deviation σ(X) when the probability distribution of a random variable X is as follows?

X	0	1	2	Total
P(X=x)	$\frac{3}{10}$	$\frac{6}{10}$	$\frac{1}{10}$	1

① E(X)=0.8, $\sigma(X)$=0.6　　② E(X)=0.6, $\sigma(X)$=0.8
③ E(X)=0.7, $\sigma(X)$=0.1　　④ E(X)=0.1, $\sigma(X)$=0.7

5.23 The number of eyes when we throw one dice is from 1 to 6. If X is a random variable which represents the number of eyes, what is the expected value of E(X)?

① 2　　② 2.5
③ 3　　④ 3.5

5.24 What is the expected value of scores if you square the number of eyes appeared when you throw one dice?

① $\frac{23}{6}$　　② $\frac{41}{6}$
③ $\frac{53}{6}$　　④ $\frac{91}{6}$

5.25 Assuming that a random variable X take three values 0, 1, and 2, and their probabilities are

$\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{6}$ respectively. What is the value of the cumulative distribution function F(x) if 1≦x<2?

① 0 ② $\frac{5}{6}$
③ $\frac{1}{3}$ ④ $\frac{2}{3}$

5.26 What is the mean and variance of the binomial distribution $B\left(100, \frac{1}{5}\right)$?

① 20, 4 ② 100, 16
③ 100, 4 ④ 20, 16

5.27 What is the variance of a random variable X which follows a binomial distribution $P(X) = {}_4C_x (0.2)^x\ (1-0.2)^{4-x}$?

① 0.16 ② 0.8
③ 0.32 ④ 0.64

5.28 There are 10 multiple choice problems and only one of the four answers is the right answer. What is the probability that four questions are correct when answering 10 questions?

① 0.0162 ② 0.0487
③ 0.1460 ④ 0.2050

5.29 According to the nature of the normal distribution, what is the probability of taking a value between $\mu \pm 3\sigma$?

① Approximately 99.73% ② Approximately 97.73%
③ Approximately 95.73% ④ Approximately 68.73%

5.30 There was a job test and the distribution of its scores is normal with the mean of 400 and standard deviation of 50. Scores between 450 and 500 are said to be suitable for the job. What percentage of scores do you think is appropriate for the job?

① 80% ② ~ 50%
③ About 30% ④ About 14%

5.31 Using the standard normal distribution table ($P(-\infty < Z < z)$) below, what is the probability of $P(-0.41 \le Z \le 2.21)$?

Z	0	1
0.4	0.6554	0.6591
2.2	0.9861	0.9864

① 0.6415 ② 0.6455
③ 0.6452 ④ 0.6418

(Answers)
5.1 ③, 5.2 ②, 5.3 ②, 5.4 ①, 5.5 ④, 5.6 ②, 5.7 ④, 5.8 ②, 5.9 ③, 5.10 ③,
5.11 ①, 5.12 ①, 5.13 ④, 5.14 ④, 5.15 ③, 5.16 ④, 5.17 ②, 5.18 ②, 5.19 ④, 5.20 ③,
5.21 ④, 5.22 ①, 5.23 ④, 5.24 ④, 5.25 ②, 5.26 ④, 5.27 ④, 5.28 ③, 5.29 ①, 5.30 ④,
5.31 ②