Statistics

Random variables & probability distribution

MATHEMATIC

❖ What is random variable

Random variable

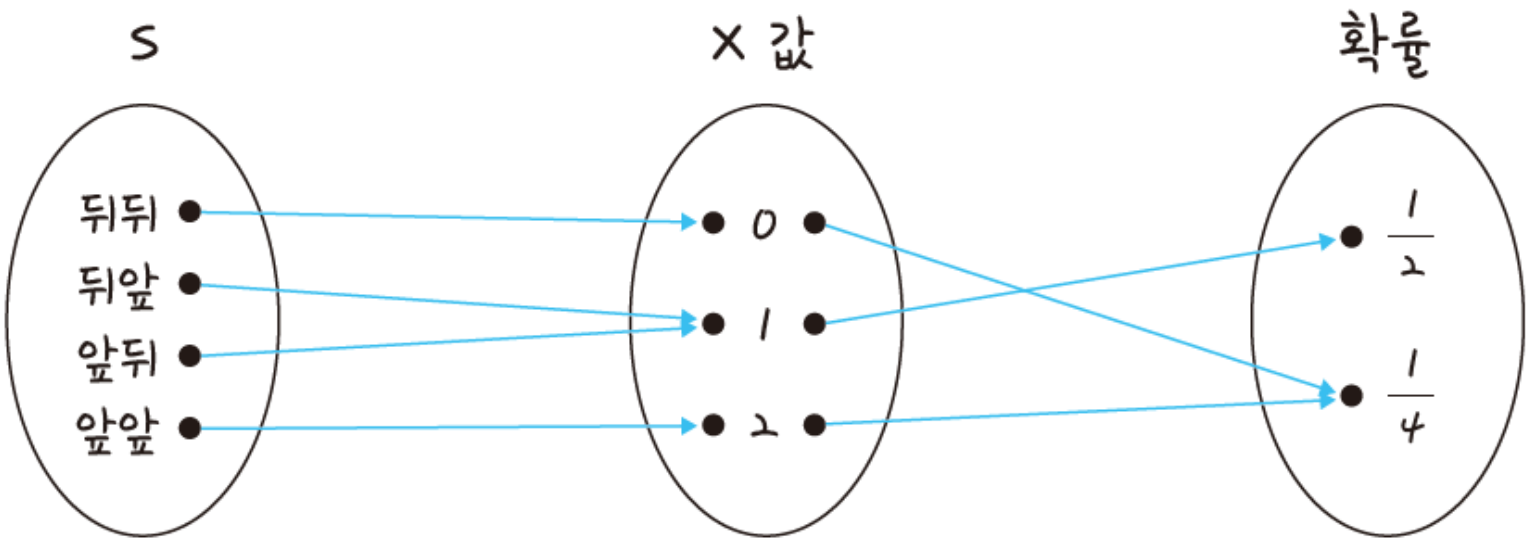
- The correspondence of one real value to each element of the sample space according to the experimental results is called ***random variable***
- For example, in an experiment in which a coin is tossed twice, if the number of heads is X, the sample space S is {back, back}, {back, front}, {front, back}, {front, front}
- Sample space : All the results of the experiment are collected and marked with S
- The number of coins that can appear on the front of the coin is 0, 1, 2 as follows
  - When X is 0: Back
  - When X is 1: front and front
  - When X is 2: front and back, front and back

Random variables & probability distribution

❖ What is random variable

- The probability that X can be 0 and 1 is $\frac{1}{4}$, respectively, and the probability that X can be 2 is $\frac{2}{4}=\frac{1}{2}$

- X is a variable with a value of one of 0, 1, and 2, and the probability corresponding to each value of X can be expressed as below



Random variables & probability distribution

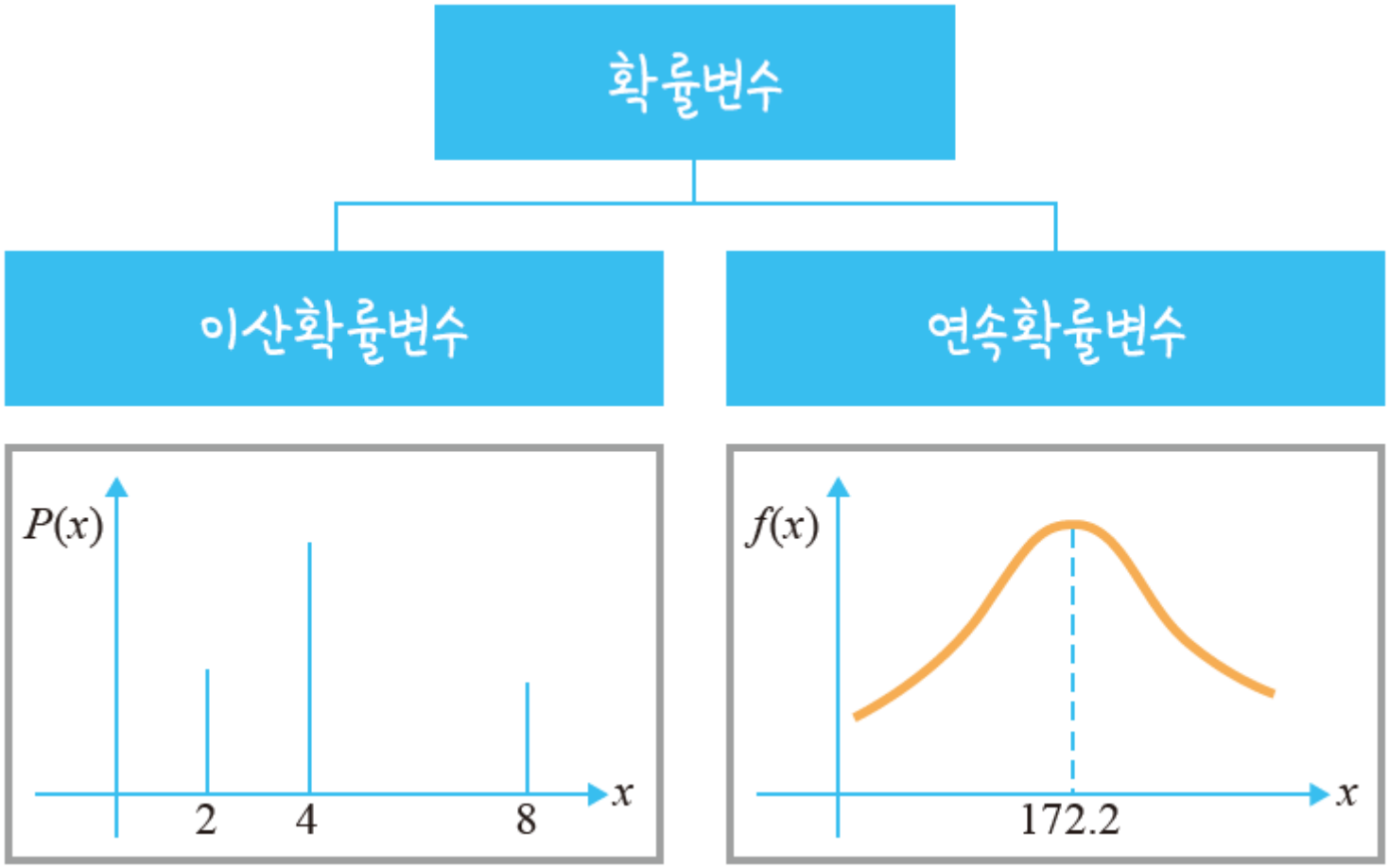
MATHEMATIC

❖ What is random variable

- Each element in the sample space corresponds to one real value: {front, front}, {front, back}, {back, front}, {back, back}
- This is called a ***function***, and when you correspond one real number to each element of the sample space, this real number is called a ***random variable***
- Probability variables can be called functions
- If a random variable is defined mathematically, the sample space is defined as a function with ***domain*** and real numbers as ***co-domain***
- The random variable is the assignment of some real value to all samples in the sample space
- Types of random variables include ***discrete random variables*** and ***continuous random variables*** as follows

Random variables & probability distribution

- Types of random variables include ***discrete random variables*** and ***continuous random variables*** as follows



Random variables & probability distribution

MATHEMATIC

❖ What is random variable

- The random variable X does not select all real numbers in any interval and is 0, 1, 2, ... This variable is a discrete random variable when only isolated values are selected
- For example, if you throw two dice, the sum of the two eyes is broken to {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}, so it can be called a discrete random variable (because there are rational and irrational numbers between the two integers, 2 and 3 cannot be called continuous)
- This variable is called a continuous random variable when the random variable *X* selects all real values in an interval
- Values such as height and weight that cannot be accurately measured are called ***continuous random variables***
- For example, if the probability variable X represents Hong Gil-dong's height and the height is close to 178, the probability according to Hong Gil-dong's height may be expressed as follows

$$P(177 < X < 179)$$

Random variables & probability distribution

MATHEMATIC

❖ What is random variable

Probability distribution

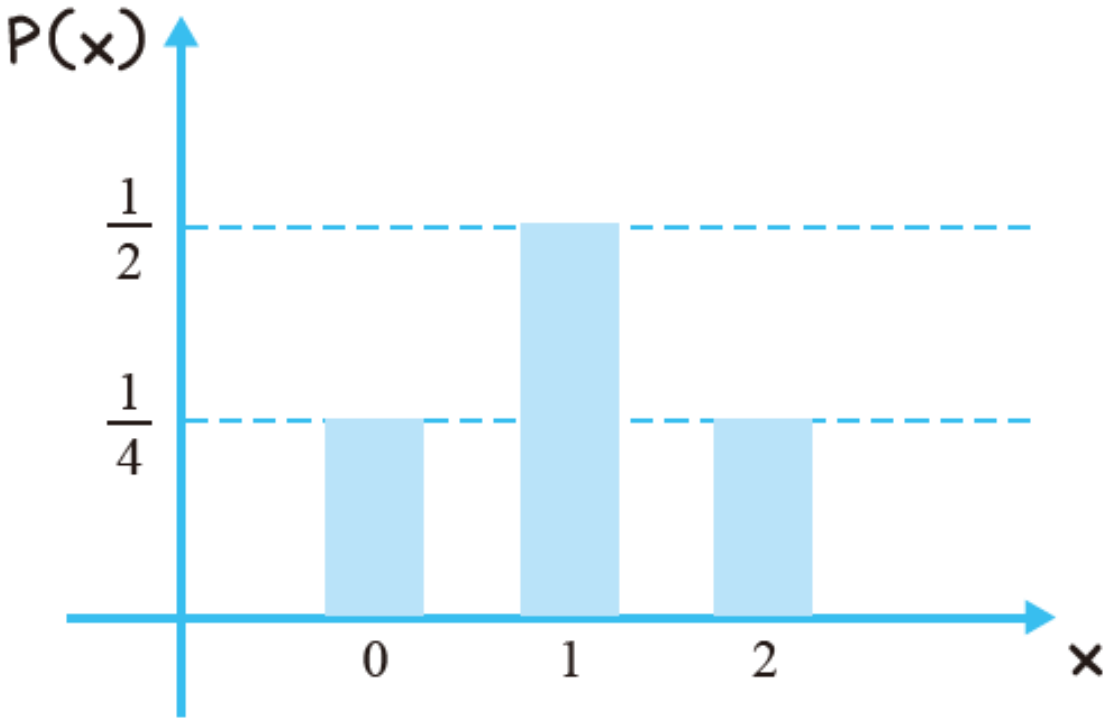
- Probability distribution is a graph of the distribution of probability values resulting from a combination of probability variables
- For example, the probability distribution of each state space value if x is the probability variable that throws a coin twice and the back side comes out

x	0	1	2
P(x)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Random variables & probability distribution

❖ **What is random variable**

- The graph of the probability distribution if x is the number of times the coin is tossed twice and the number of tails comes out



Random variables & probability distribution

MATHEMATIC

❖ What is random variable

Probability function

- A probability function is a function representing the probability distribution of discrete random variables, also known as ***a probability mass function (PMF)***
- When a random variable has a finite number of values or is a discrete random variable that can be counted such as a natural number, the function representing the probability of that discontinuous value is a probability function
- The probability that an X variable takes a particular value of x is called a probability function and is expressed as f(x)

$$f(x) = P[X = x]$$

Random variables & probability distribution

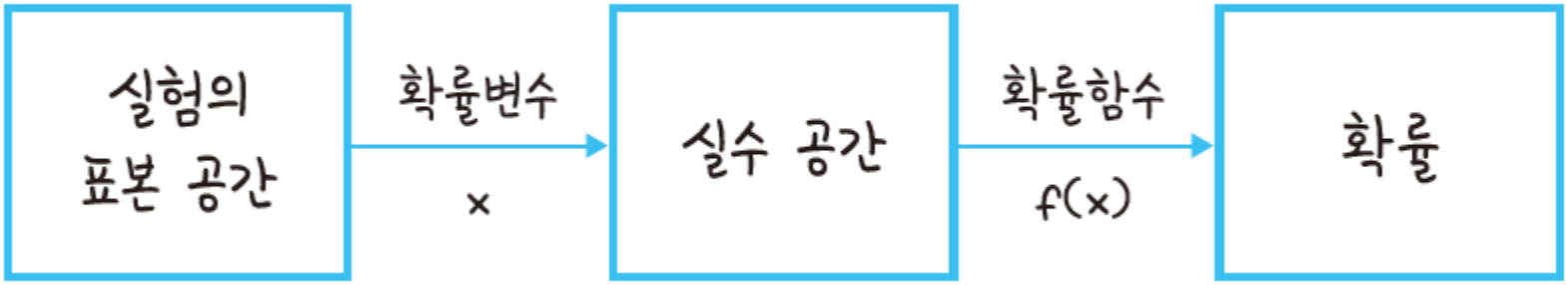
❖ What is random variable

- For example, if the number of heads in an experiment in which a coin is thrown twice is called a random variable X, the values that the random variable X can take are 0, 1, and 2, and each probability is as follows

(1) When X is 0 $P(X=0)=\frac{1}{4}$ (back/back)

(2) When X is 1 $P(X=1)=\frac{2}{4}=\frac{1}{2}$ (front/back and back/front)

(3) When X is 2 $P(X=2)=\frac{1}{4}$ (front/front)



Random variables & probability distribution

MATHEMATIC

❖ What is random variable

- A ***probability function*** is a function representing the probability that a random variable will occur, so if you know the probability function of a specific random variable, you can predict the probability that a specific event will occur, which is necessary in statistics
- The stats sub-package provided by Python's SciPy library provides a variety of features to analyze probability distributions

In [9]:

```
# stats 서브패키지를 호출합니다
# 한글 깨짐을 방지하는 코드
import matplotlib as mpl
import matplotlib.pylab as plt
from matplotlib import font_manager

font_fname = 'C:/Windows/Fonts/malgun.ttf'
font_family = font_manager.FontProperties(fname=font_fname).get_name()

plt.rcParams["font.family"] = font_family
```

Random variables & probability distribution

MATHEMATIC

❖ What is random variable

In [10]:

```
from scipy import sp
import seaborn as sns
import numpy as np
import matplotlib as mpl
import matplotlib.pylab as plt
%matplotlib inline
# 확률질량함수는 SciPy의 stats 서브패키지에 binom 클래스로 구현합니다
N = 10   # 전체 시도 횟수
mu = 0.6 # 베르누이 확률분포의 기댓값
rv = sp.stats.binom(N, mu)
xx = np.arange(N + 1)
```

Random variables & probability distribution

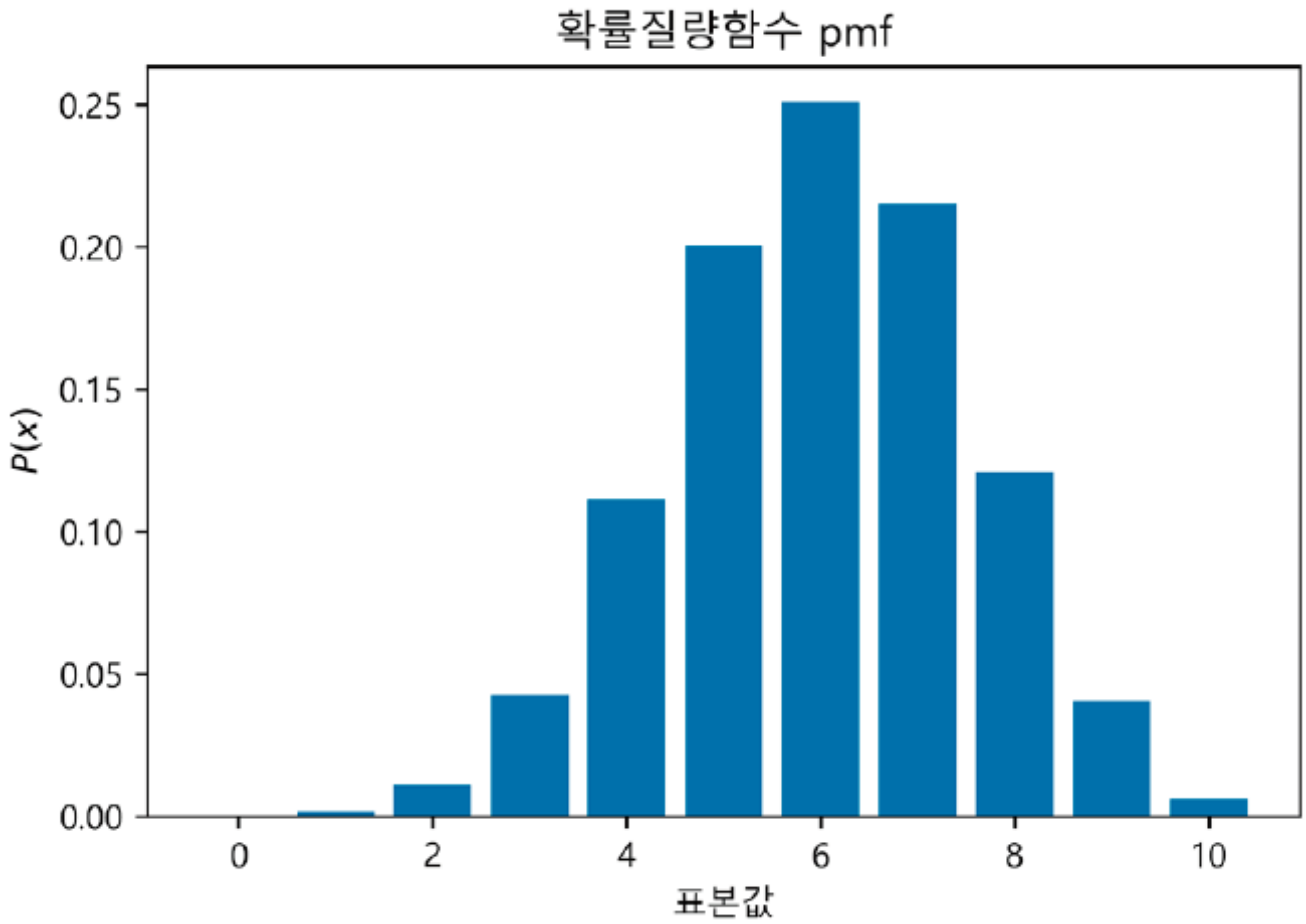
MATHEMATIC

❖ What is random variable

```
# 그래프를 표현할 때는 matplotlib을 사용합니다
plt.bar(xx, rv.pmf(xx), align="center") # 확률질량함수 기능을 갖는 pmf를
사용합니다
plt.xlabel("표본값")
plt.ylabel("$P(x)$")
plt.title("확률질량함수 pmf")
plt.show()
```

Random variables & probability distribution

❖ **What is random variable**



Random variables & probability distribution

MATHEMATIC

❖ What is random variable

Probability density function (PDF)

- Continuous random variables are impossible to express the distribution because the values that the random variables can take are continuous and infinite
- If a particular random variable corresponds to a particular random value, such as a discrete type, the probability that a particular random variable will be included in a particular interval becomes infinite (because there are infinite values that the random variable can take)
- ***Probability density function (PDF)*** is required to solve this problem of continuous random variables
- A probability density function is a function that calculates the probability of belonging to a specific interval, and in the graph represented by the function, the function becomes 'area of a specific interval = probability of belonging to a specific interval'

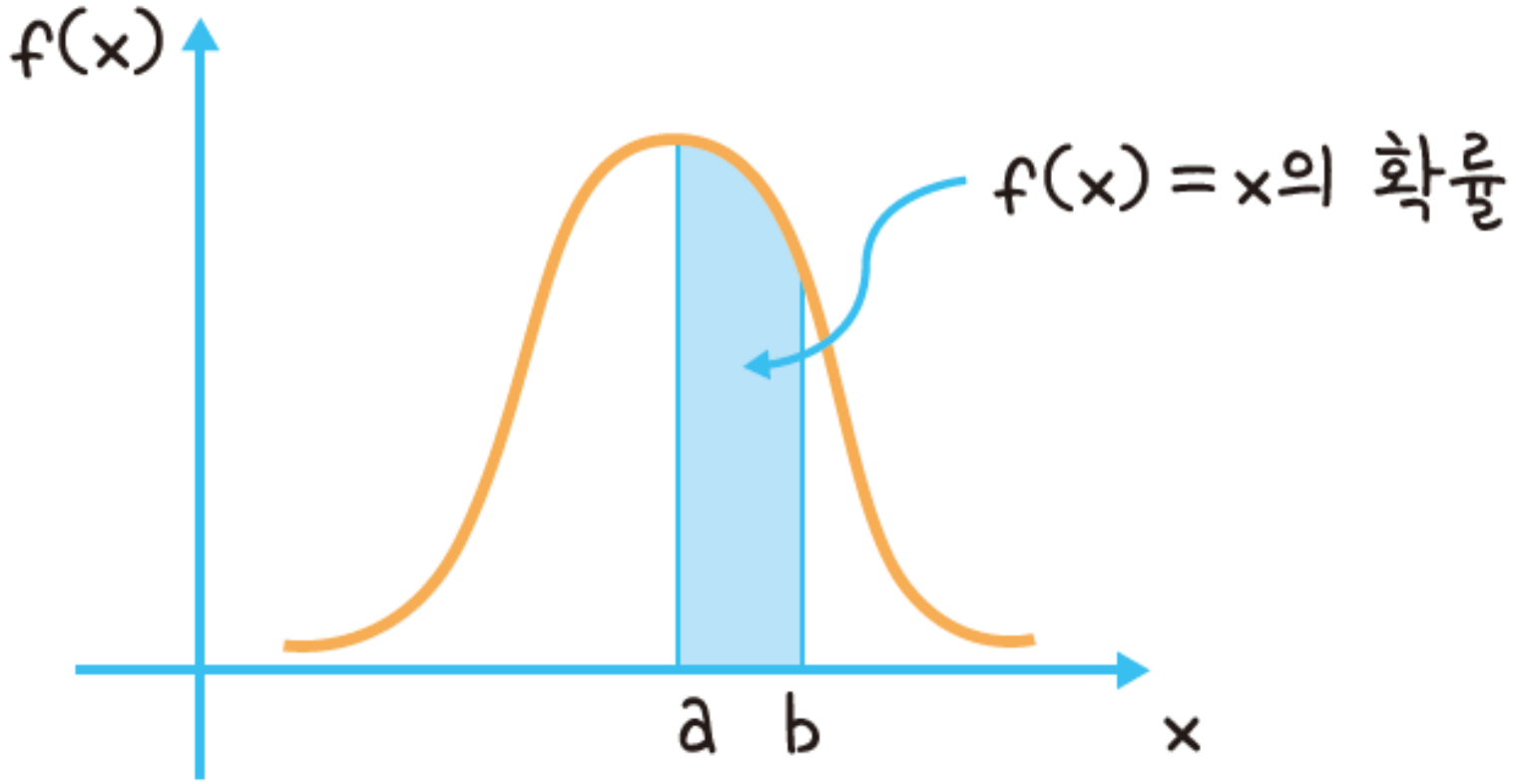
Random variables & probability distribution

MATHEMATIC

❖ What is random variable

- The conditions of the probability density function are as follows
  - (1) f (x) ≥ 0 for all x values
    - For all real values of x, the probability density function is greater than or equal to 0
  - (2) The probabilities of all possible values of x are obtained by integral $\int_{-\infty}^{\infty} f(x)dx$,and this value is always 1
  - (3) The probability of an interval (a, b) is $\int_{a}^{b} f(x)dx$
    - The probability of x for an interval (a, b) is the size of the area created by the probability density function f(x) in that interval

Random variables & probability distribution



Random variables & probability distribution

MATHEMATIC

❖ What is random variable

- Even with Python, the probability density function can be obtained as follows

```
In [11]:
# SciPy 라이브러리에서 제공하는 sp 서브패키지를 호출합니다
from scipy import sp
import seaborn as sns
import numpy as np
import matplotlib as mpl
import matplotlib.pylab as plt
%matplotlib inline
```

Random variables & probability distribution

MATHEMATIC

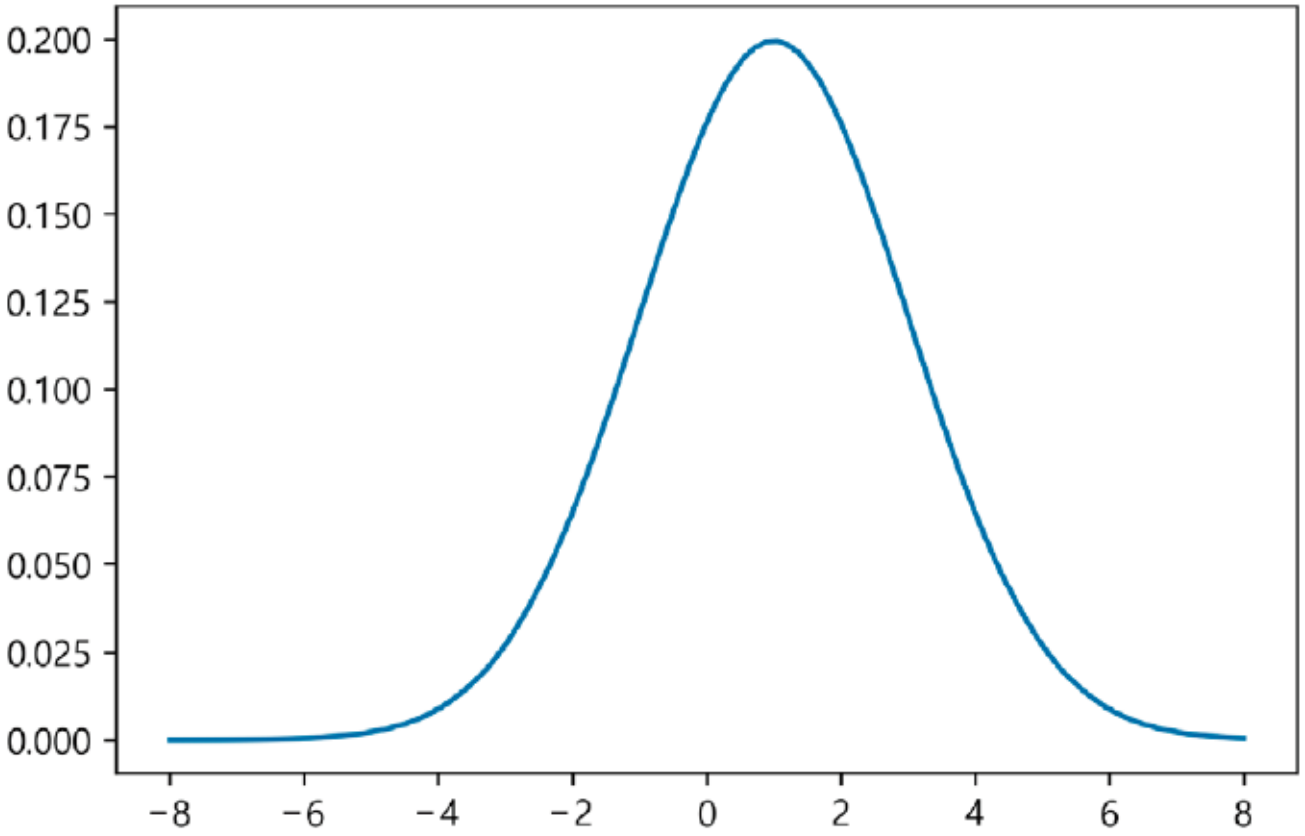
❖ What is random variable

```
# 기댓값이 1이고 표준편차가 2인 정규분포 객체를 생성합니다
rv = sp.stats.norm()
rv = sp.stats.norm(loc=1, scale=2)

# 확률분포 객체의 메서드 중 확률밀도함수 기능을 갖는 pdf를 사용합니다
x = np.linspace(-8, 8, 100)
pdf = rv.pdf(x)
plt.plot(x, pdf)
plt.show()
```

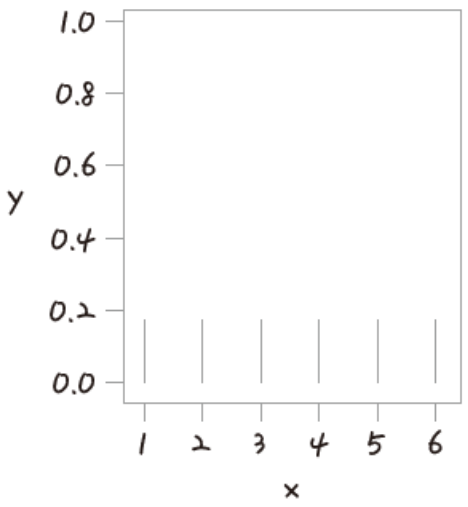
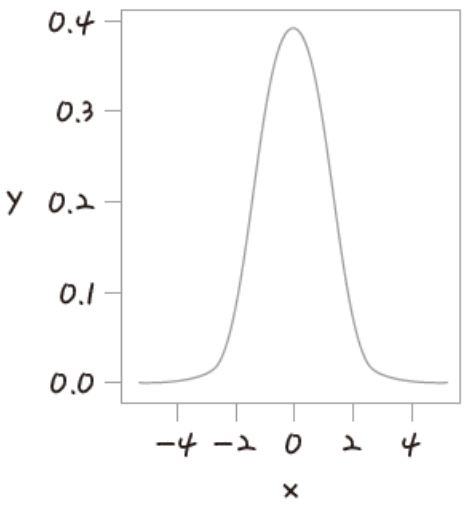
Random variables & probability distribution

MATHEMATIC

- **What is random variable**



Random variables & probability distribution

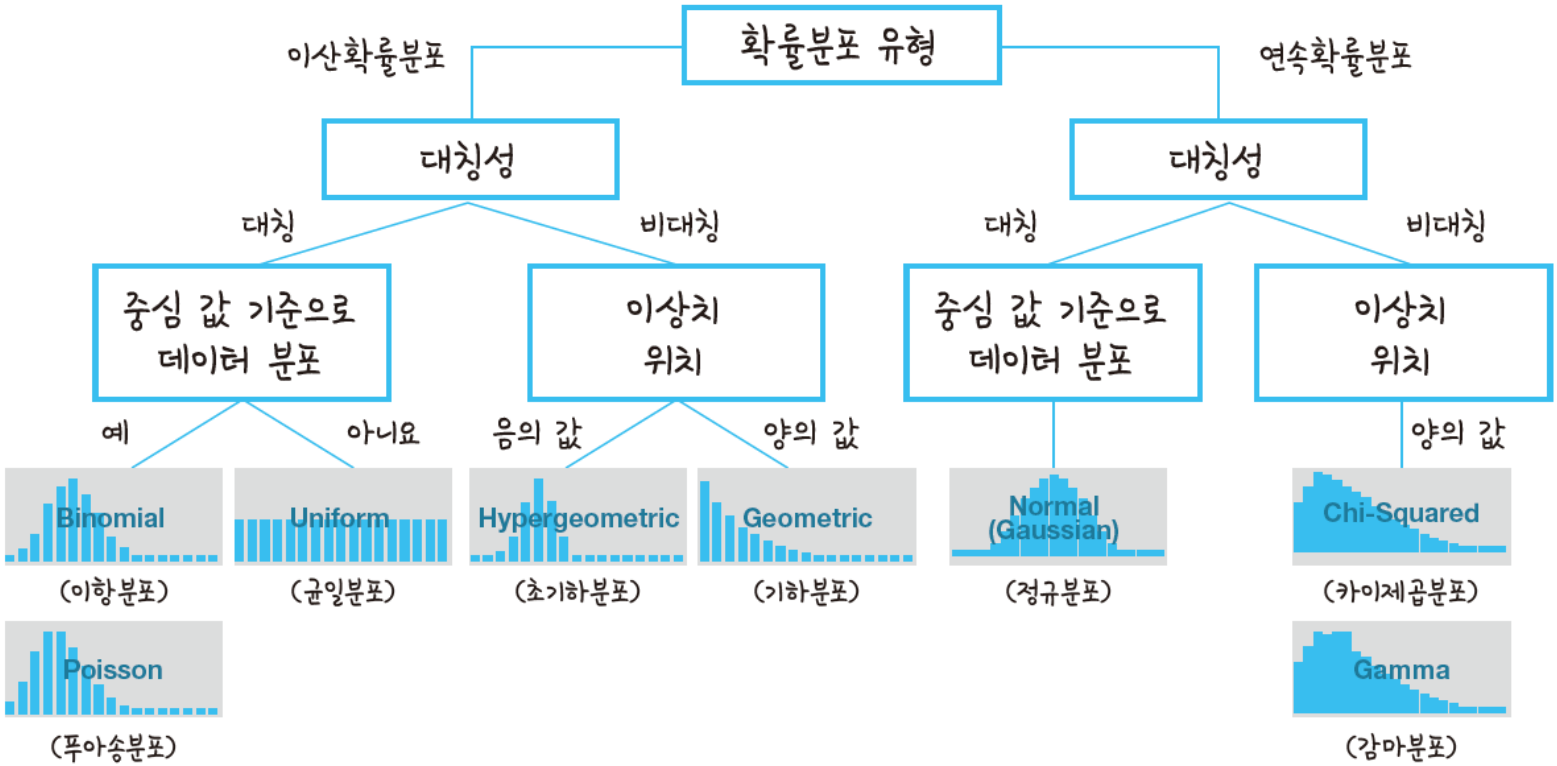
구분	확률질량함수	확률밀도함수
분류	이산형 확률변수	연속형 확률변수
함수	각 이산점의 확률 크기를 표현하는 함수	연속형 데이터의 확률을 표현하는 함수
변수량	유한	무한
그래프		

Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

- The probability distribution has various types according to the discrete and continuous probability distributions, and can be subdivided as follows according to the data distribution or outlier value

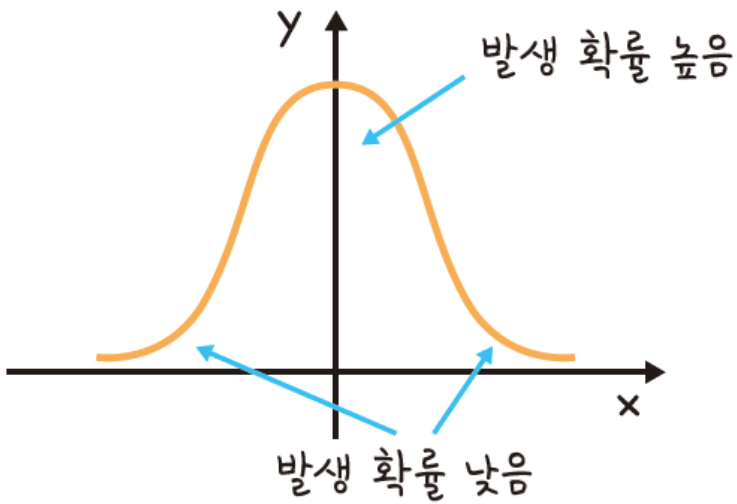


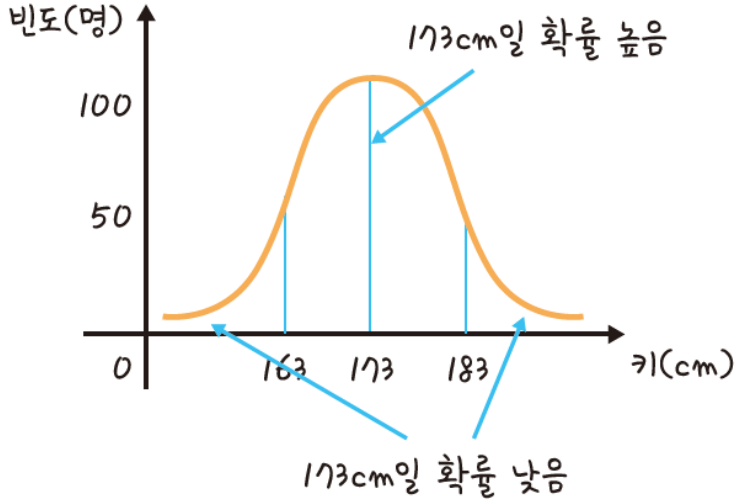
Random variables & probability distribution

❖ Types of probability distributions

Normal distribution and standard normal distribution

- Normal distribution is the most used probability distribution in statistics
- Used to understand phenomena such as economic, business, and social issues
- The normal distribution is also called a ***Gaussian distribution***
- It refers to a distribution that is closer to the mean, more likely it is to occur, and less likely to occur as it is further away from the mean

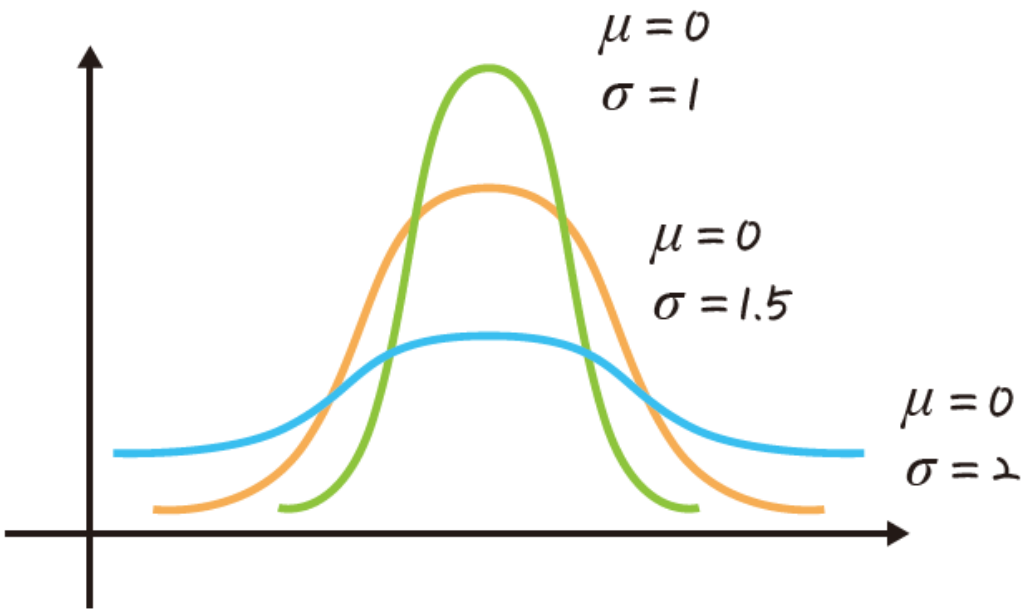




Random variables & probability distribution

❖ Types of probability distributions

- Mathematically, it is a ***continuous probability distribution with a mean of μ and a standard deviation of σ***.
- The larger the standard deviation σ, the more it becomes a spread bell shape, and the smaller the σ, the more concentrated it becomes a pointed bell shape

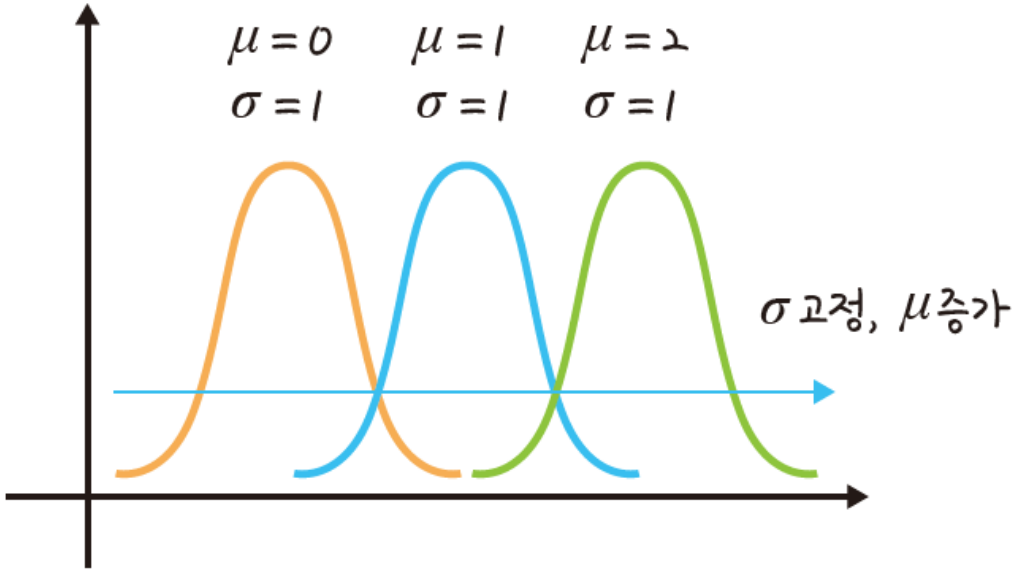


Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

- The properties of the normal distribution curve are as follows
  - (1) Symmetrical to straight line x = μ (average)
  - (2) Total area is 1
  - (3) When μ is constant, the width of the graph widens as σ increases
  - (4) When σ is constant, the larger μ, the more to the right



Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

- It is difficult to compare data because the mean and standard deviation of each group are different
- For example, if the result of the class A math test is 80 points, the standard deviation is 40 points, and the result of the class B test is 60 points, and the standard deviation is 12 points, which class is better?
- It is difficult to judge intuitively because the distribution of math score data in Class A and B is different
- ***Normal distribution*** must be standardized to compare groups with normal distributions with different parameter values (***mean, standard deviation***), which is called ***standard normal distribution***

Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

Standard normal distribution

- Standardized by making the mean of the normal distribution '0' and standard deviation ‘1’
- It's simple to make the mean 0 and the standard deviation
- It can subtract the average (μ) of the entire data from the random variable X of the collected individual data and divide it by the standard deviation (σ)

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

Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

- Since the individual data is subtracted by the average of the total data, the average of the individual data is 0 again
- This means that it's moved horizontally to 0
- These standardized individual data are called the standardized score (z-score), and the standardized score becomes a normal distributed random variable with a mean of 0 and a standard deviation of 1

Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

- Python can also express normal distributions as follows

In [12]:

```
# NumPy와 matplotlib 라이브러리를 호출합니다
import numpy as np
import matplotlib.pyplot as plt

# 평균 및 표준편차를 정의합니다
mu, sigma = 0, 0.1

# np.random.normal 함수를 사용해서 평균 0, 표준편차 0.1인 샘플 1000개를 추출합니다
s = np.random.normal(mu, sigma, 1000)
```

Random variables & probability distribution

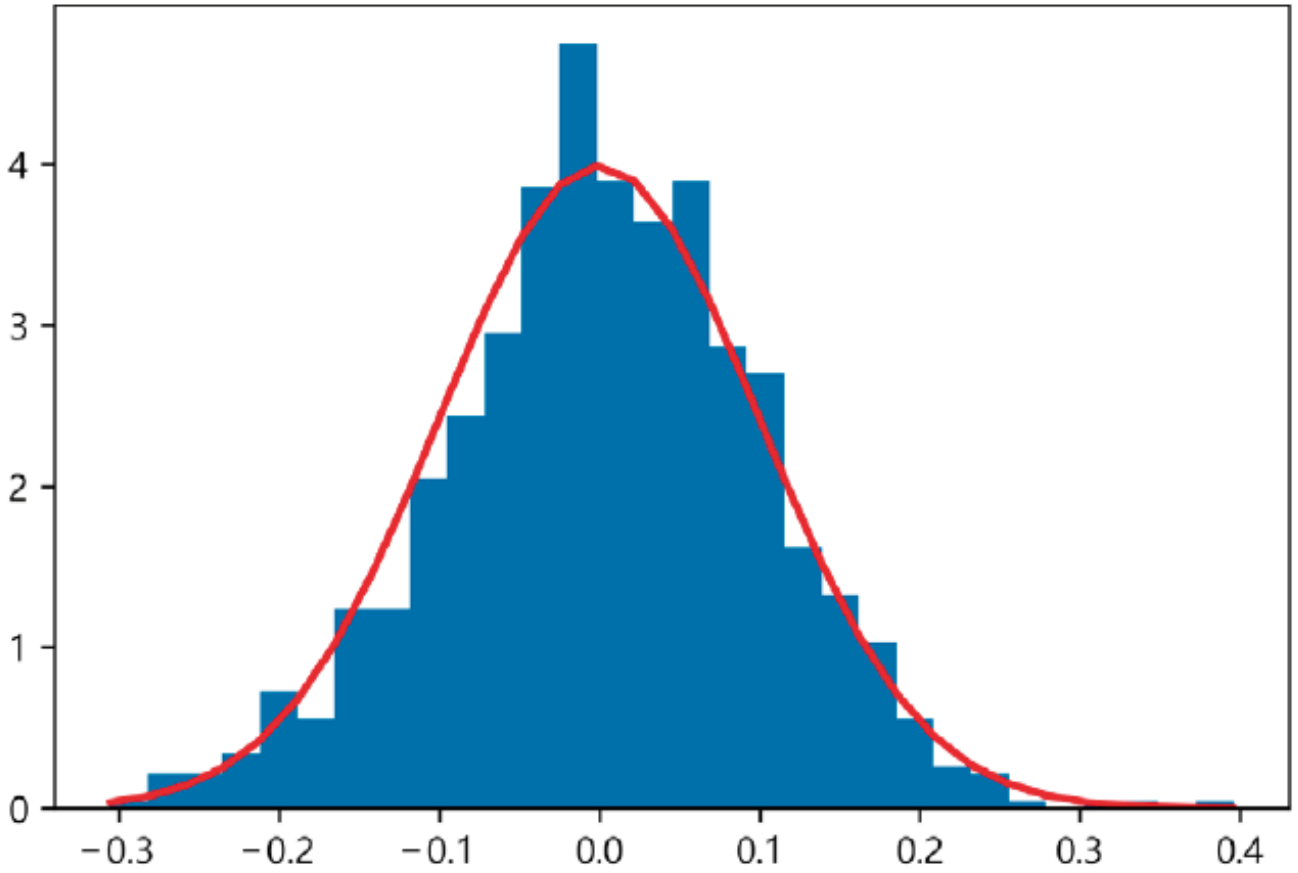
MATHEMATIC

❖ **Types of probability distributions**

```
# 샘플들의 histogram을 출력합니다
# (s: 배열 혹은 배열들로 구성된 시퀀스,
# 30: 해당 막대의 영역(bin)을 얼마나 채우는지 결정하는 변수)
count, bins, ignored = plt.hist(s, 30, normed=True)

# 샘플들을 이용해서 정규분포의 모양으로 출력합니다
# (plot(x축 데이터, y축 데이터)꼴로 사용)
plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *
    np.exp( - (bins - mu)**2 / (2 * sigma**2) ), linewidth=2, color='r')
plt.show()
```

Random variables & probability distribution

❖ **Types of probability distributions**

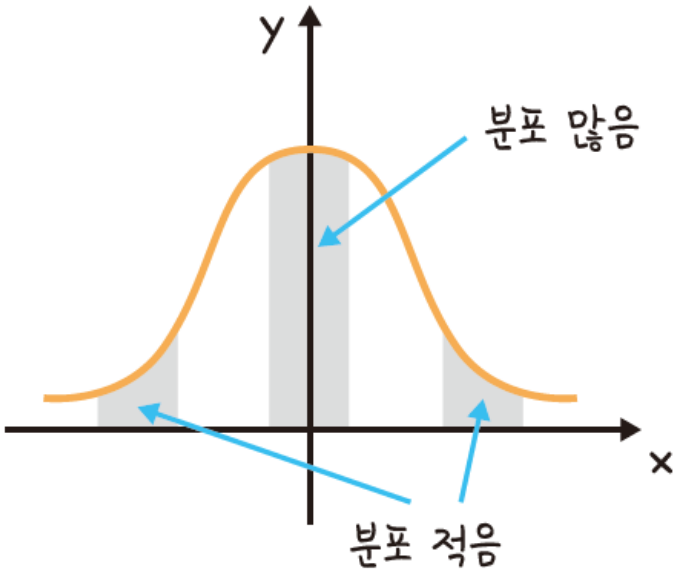


Random variables & probability distribution

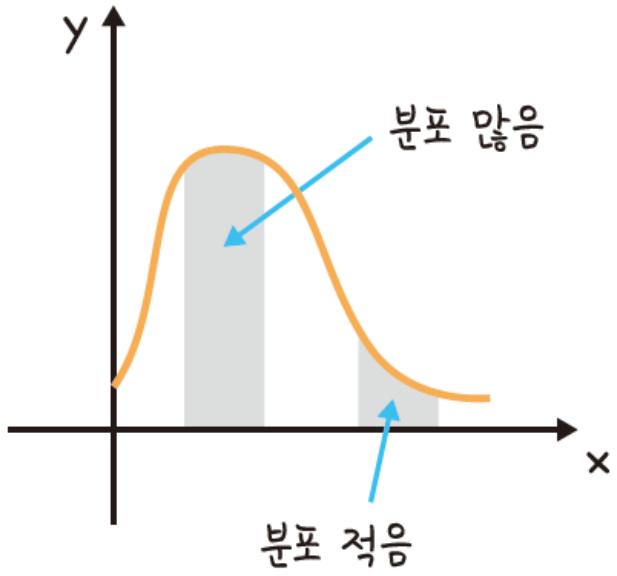
❖ Types of probability distributions

Chi-squared distribution

- Chi-squared distribution (X^2 distribution) is often used to estimate and test the variance of a group



① 정규분포



② 카이제곱분포

❖ Types of probability distributions

- The chi-square distribution is more distributed as it approaches 0, and the distribution decreases as it moves away from 0
- Graphs like this appear because the bias of data or groups is not very large
- For example, the average height of adult men in Korea is 173cm, which means that many people are close to the average height, such as 169.8cm and 181.2cm, but not many people are 150.2cm and 193.3cm tall
- The chi-square distribution has a higher distribution near 0, and the further away from 0, the lower the distribution

Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

Binomial distribution

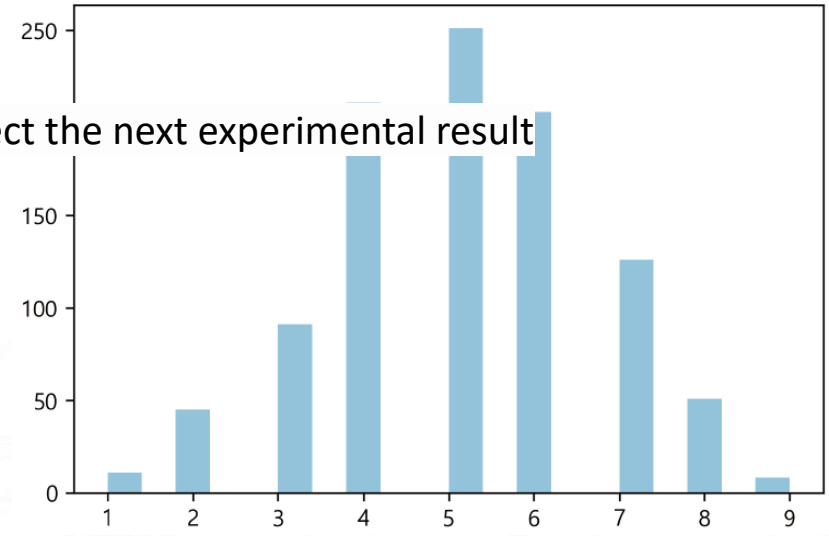
- Since the binary distribution is based on the Bernoulli distribution
- Bernoulli experiment is a probability experiment with two outcomes, such as yes or no
- The success rate is that the experimental result is success, and the probability of success is called *p*
- For example, tossing a coin and getting a head or a back is also Bernoulli's experiment
- Bernoulli trial is an independent repetition of the Bernoulli experiment

Random variables & probability distribution

MATHEMATIC

❖ Types of probability distributions

- The following are the conditions of the Bernoulli experiment
  - Each experimental result is divided into two mutually exclusive events (success or failure)
  - The probability that the result will succeed in each experiment is expressed as p = P(S), and the probability that the result will fail is expressed as q = P(F) = 1 – p
  - p + q = 1
  - Each experiment is independent
  - One experimental result does not affect the next experimental result



Statistical inference

MATHEMATIC

❖ Mean, median, mode

Representative value

- A representative value is a representative value that best describes the data
- When you have tens thousands of pieces of data, it's not efficient to check them one by one
- What data means is important, and it is a representative value that can effectively express it
- The most frequently used representative values are ***mean, median,*** and ***mode***

❖ Mean, median, mode

Mean

- The mean is a mathematical measure of the central tendency in a group, which is the value of the sample added together and divided by the number of samples

$$\text{평균} = \frac{\text{표본의 총합}}{\text{표본의 개수}}$$

- For example, if you have data 1, 1, 3, 5, 6, 7, 8, 9, 10, mathematically, you can calculate it

$$\frac{1+1+3+5+6+7+8+9+10}{9} = 5.6$$

- The mean has the advantage of being able to use all data values, but the disadvantage is that using extreme data values can distort representative values

❖ Mean, median, mode

Median

- Median refers to the most central value when the given values are sorted by size
- If the total number of data is odd, the value at the center is the median; if it is even, the mean of the two values at the center is the median

전체 데이터 개수(n)가 홀수일 때: $\frac{n+1}{2}$ 번째 값

전체 데이터 개수(n)가 짝수일 때: $\frac{n}{2}$, $(\frac{n}{2}+1)$번째 값들의 평균

2, 3, 4, 5, 6, 6, 7, 7, 8,

↑

중앙값

2, 3, 5, 6, 7, 7, 8, 9, 10, 11

↑

중앙값 $\frac{(7+7)}{2}=7$

Statistical inference

MATHEMATIC

❖ Mean, median, mode

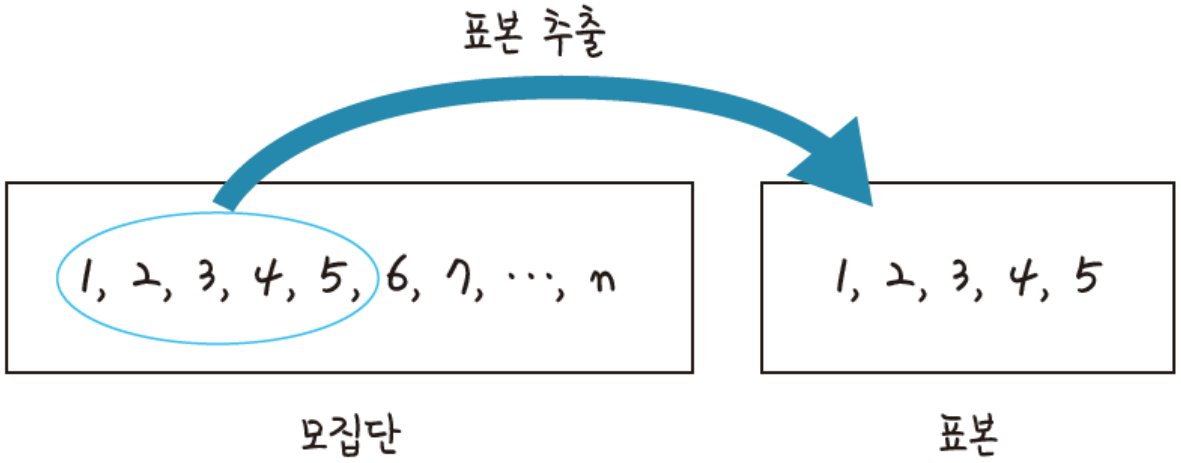
Mode

- The mode is the most frequently observed number, i.e., the most frequently seen value of a given value
- For example, when 1, 1, 2, 3, 4, 6, 6, 6, 6, 6, 6 is rolled ten times, the most frequently appeared '6' becomes the most frequent value
- The advantage of the mode is that it is useful to find the most common value
- Data can be used even if it is not a number
- The relationship between the number of favorite numbers or idol groups is often used for meaningless qualitative data

❖ Population distribution and random sample

Population and parameter

- A population is any object that is subject to any statistical experiment
- The probability distribution formed by the data constituting the population is a **population distribution**
- The way to know the parameters (characteristics) of a population is to estimate the parameters of the population by sampling a sample



❖ Regression analysis

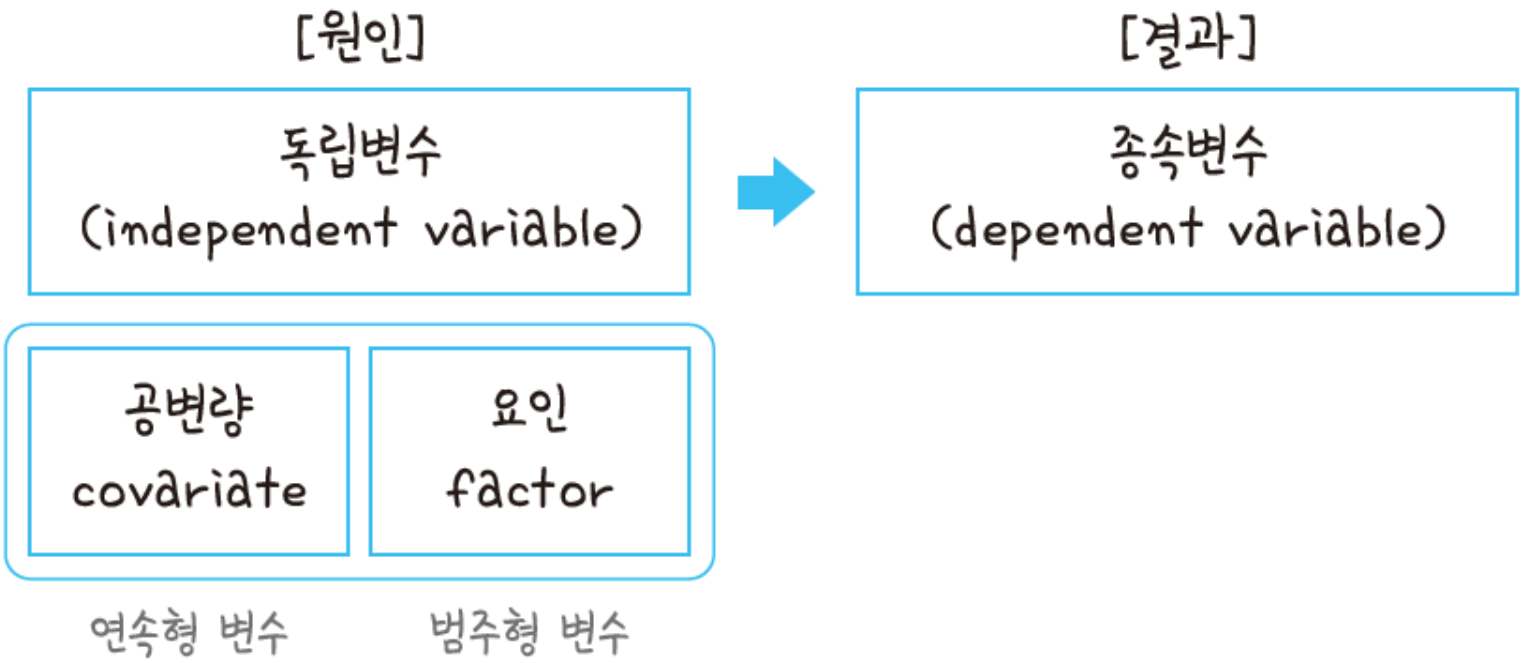
Independent and dependent variables

- Assuming that the value of the function y also changes as the value of the x variable changes, the expression y = f(x) is established
- x is called the independent variable, and y is called the dependent variable
- The independent variable is a variable that changes according to the researcher's intention, and the dependent variable is a variable that the researcher wants to know how it changes according to the independent variable that changes

Statistical inference

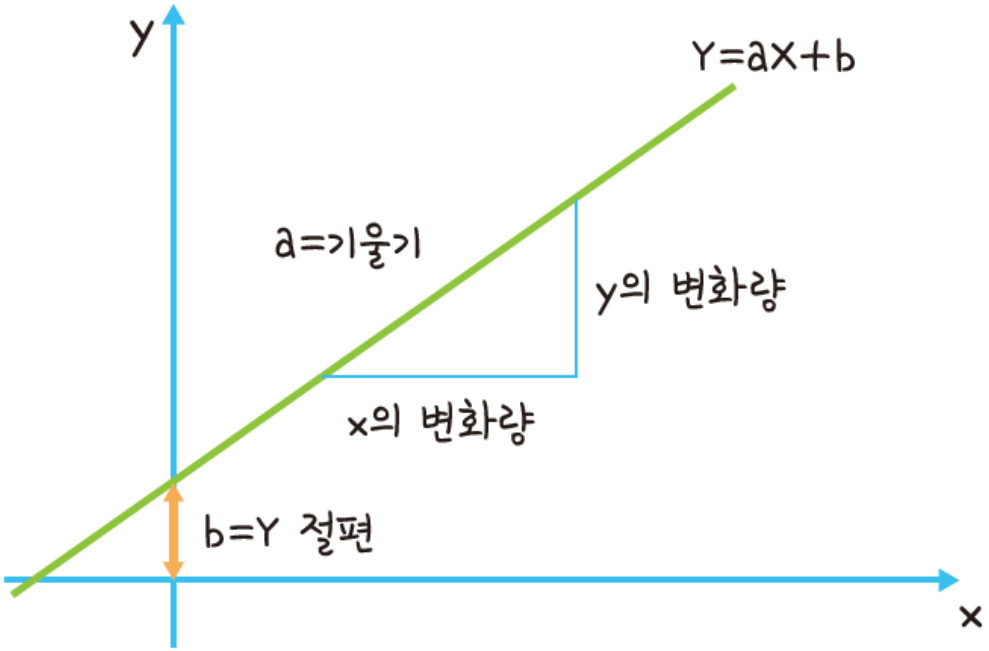
❖ Regression analysis

- As such, independent and dependent variables have a causal relationship



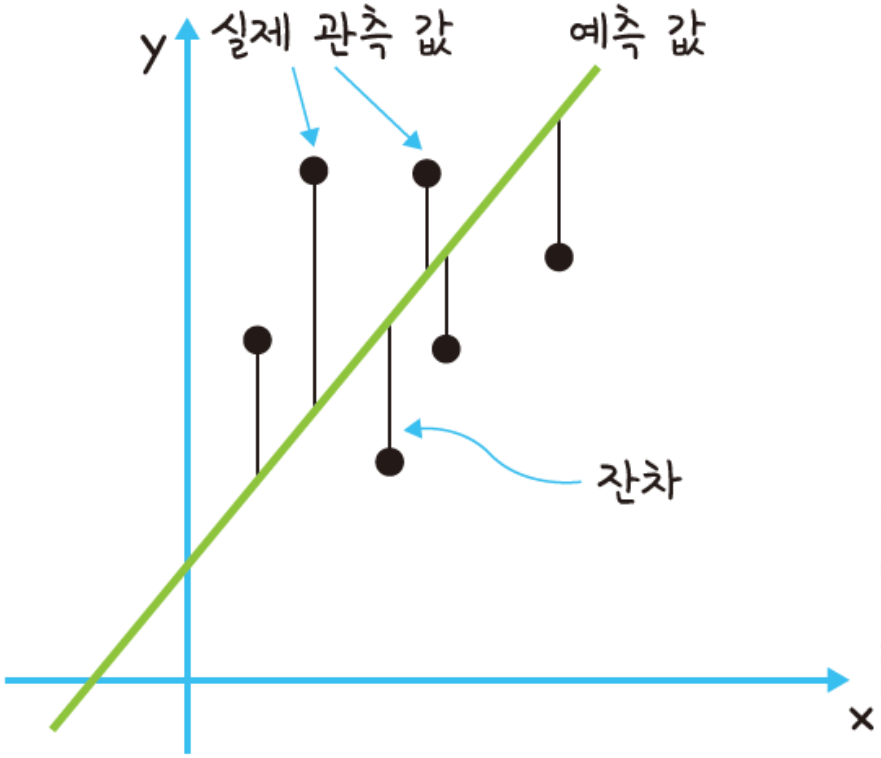
❖ Regression analysis

- Regression analysis has the following characteristics
  - ***Regression***: It means 'regression to the mean' and means that the relationship between two variables returns to the mean of the generalized linear relationship
  - ***Linearity***: Two variable relationships can be described as one straight line (primary equation: Y = aX + b)



❖ Regression analysis

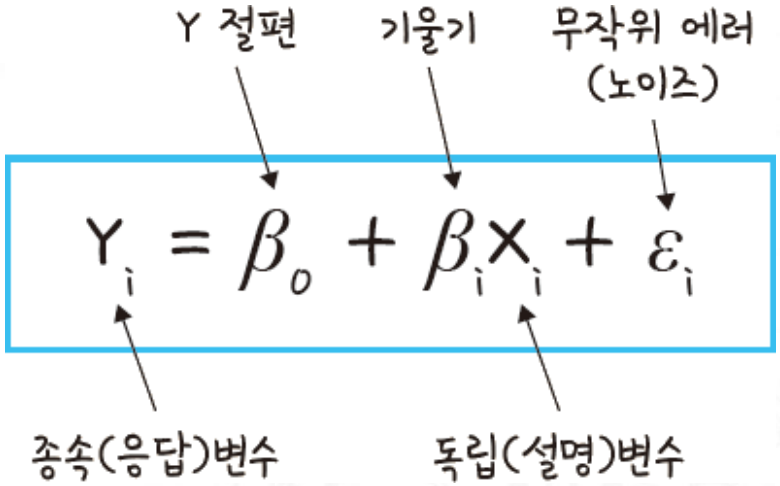
- ***Linear relationship***: Estimate parameters using a linear regression equation, and use the least squares method as a straight line that minimizes the sum of the difference (residuals) between the predicted value and the actual observed value



Statistical inference

❖ Regression analysis

- Although the data is mapped exactly to a straight line, it is almost impossible in reality to obtain such a straight line
- Used by adding random errors (noise)
- In the previous equation, the relationship between β_0 and β_i that satisfies the condition that the ε mean becomes 0 is a linear regression equation, and this leads to a linear model (called regression because the values of the parameters (β_0, β_i) are narrowed to minimize the ε mean)

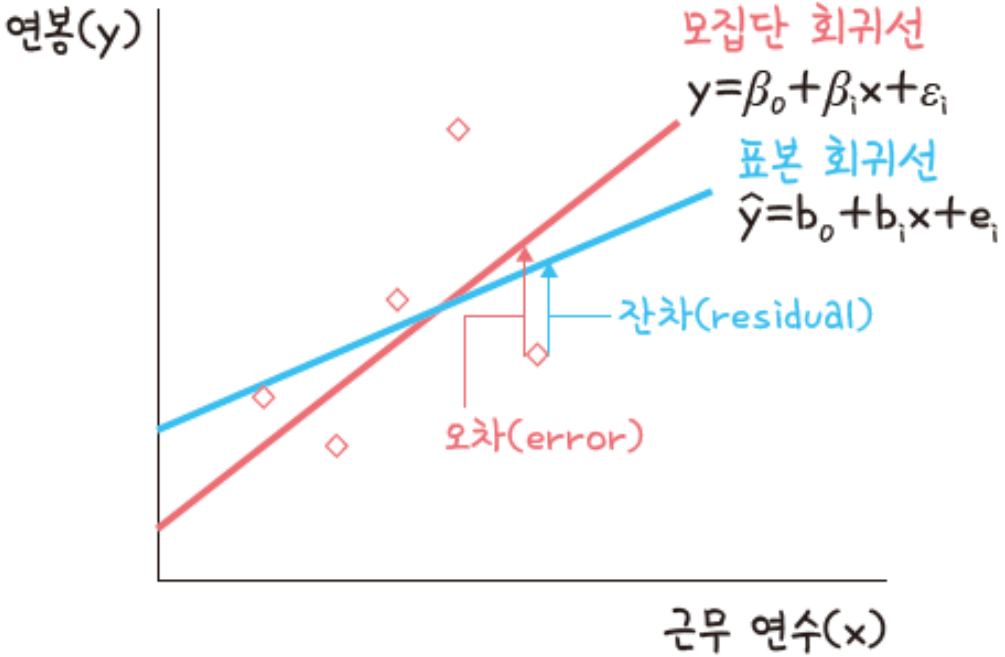


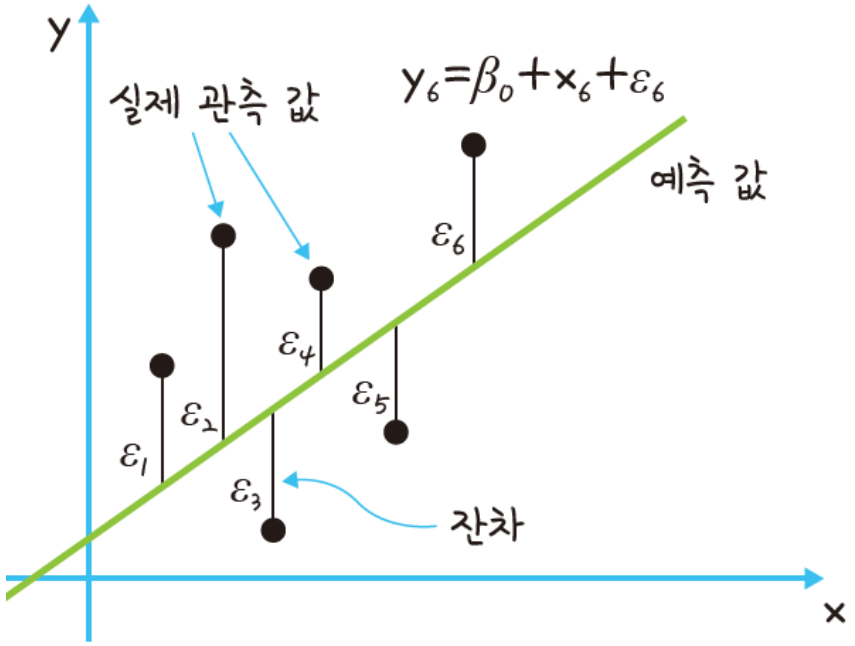
❖ Regression analysis

Residual and error

- The regression equation refers to a straight line representing a population, and since very few have a population, we infer the regression equation of the sample group that can represent the population
- The concepts that came out of this are residual and error
  - ***Residual*** = Predicted value from regression expression of sample group – actual observation value
  - ***Error*** = Predicted value from regression expression of population – actual observation value

Statistical inference



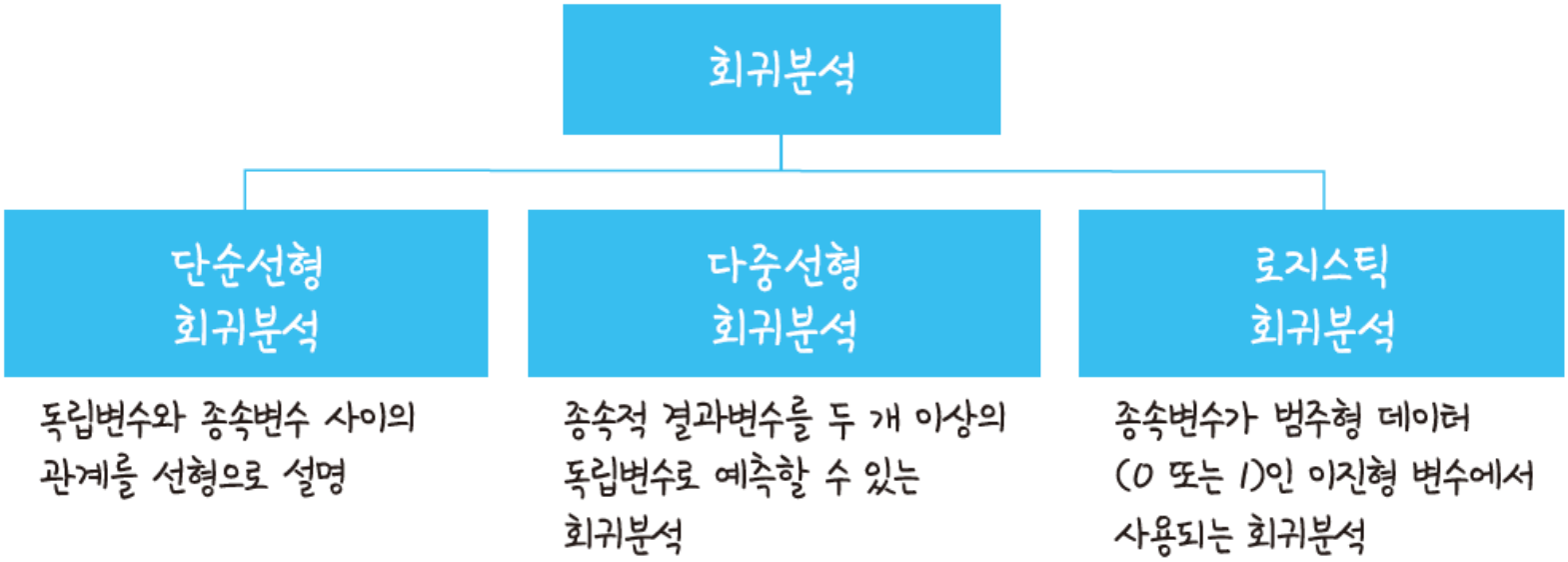


Statistical inference

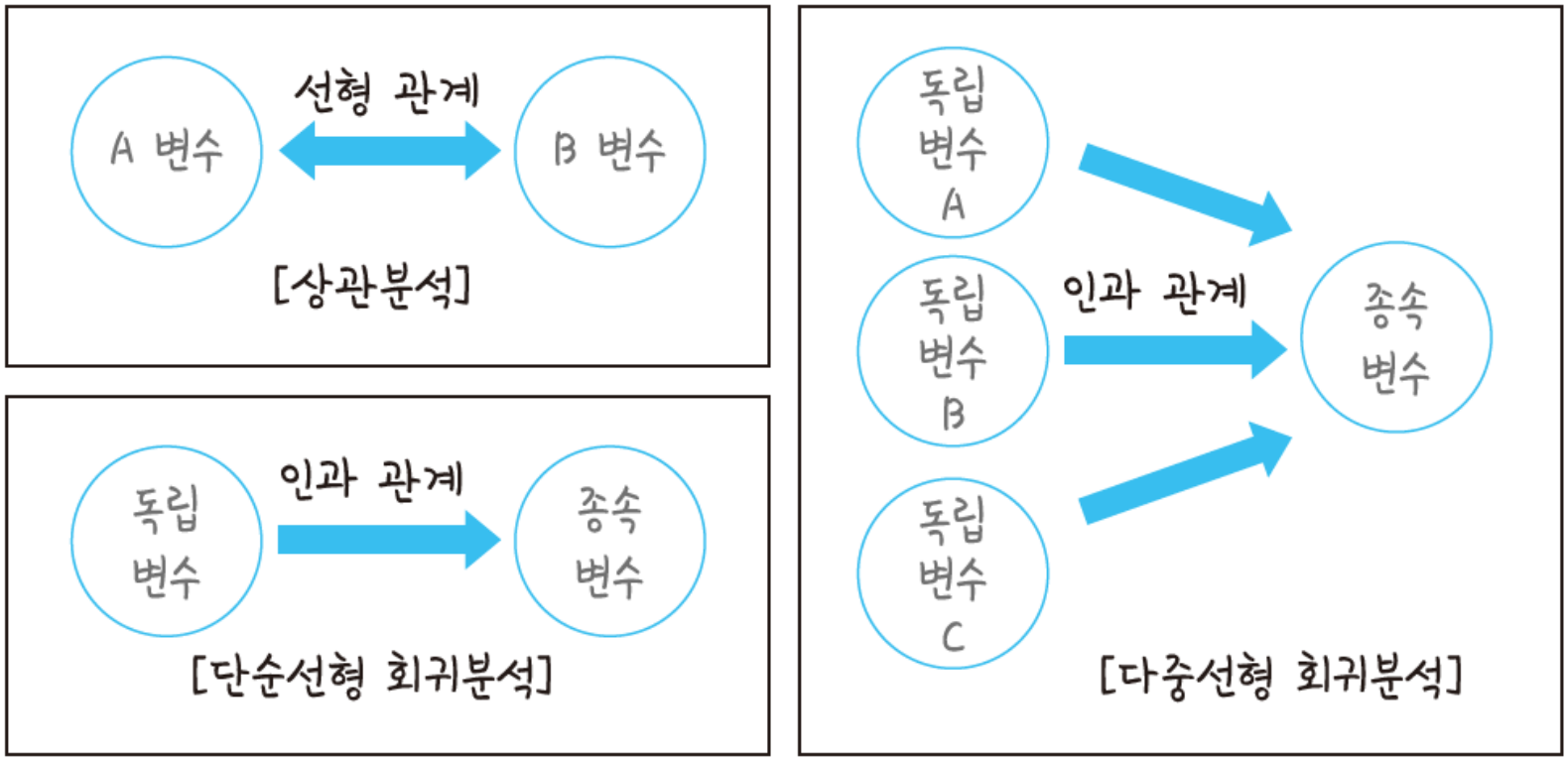
❖ Regression analysis

Types of regression analysis

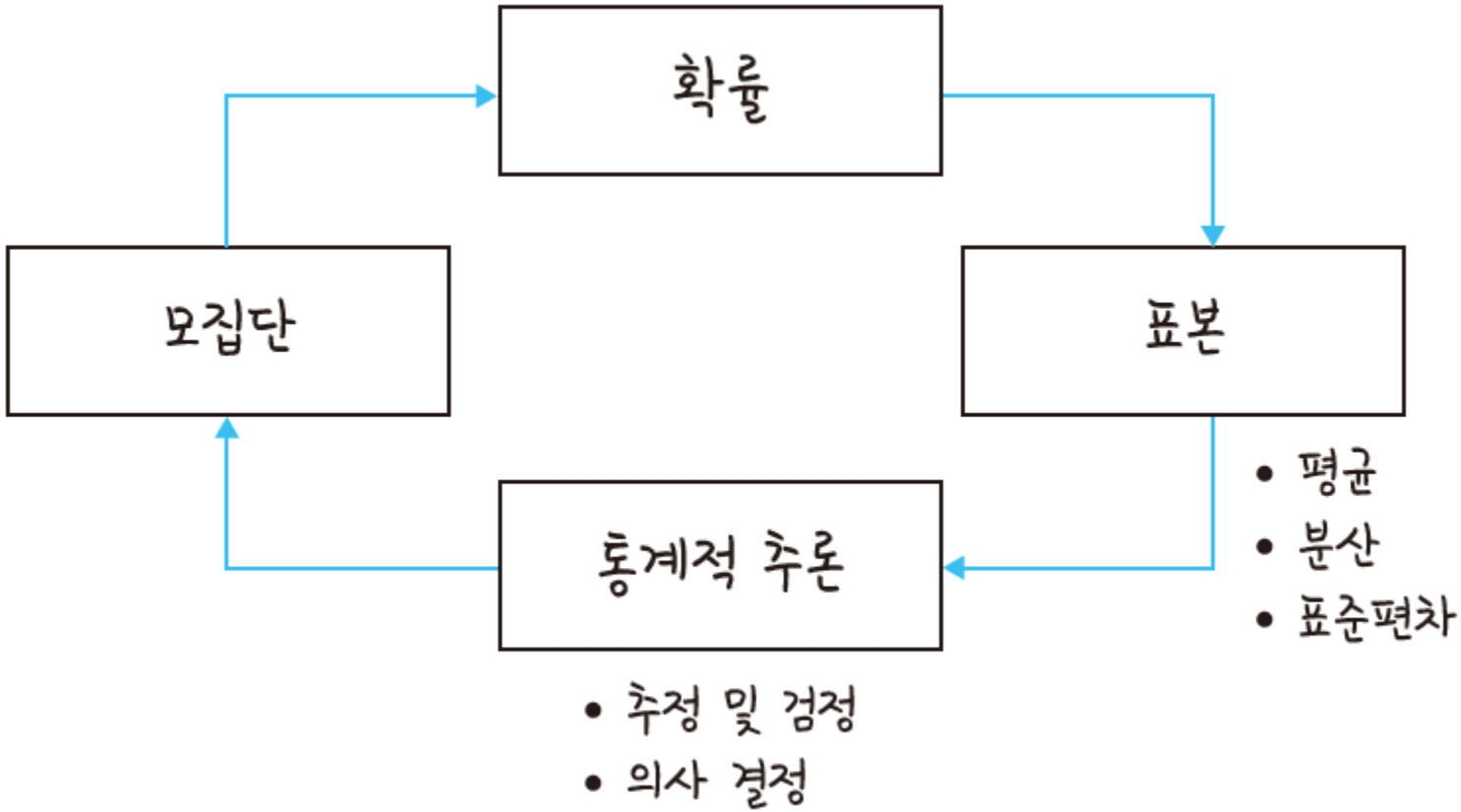
- Simple linear regression, multiple linear regression, and logistic regression



Statistical inference



Statistical inference



Statistical test

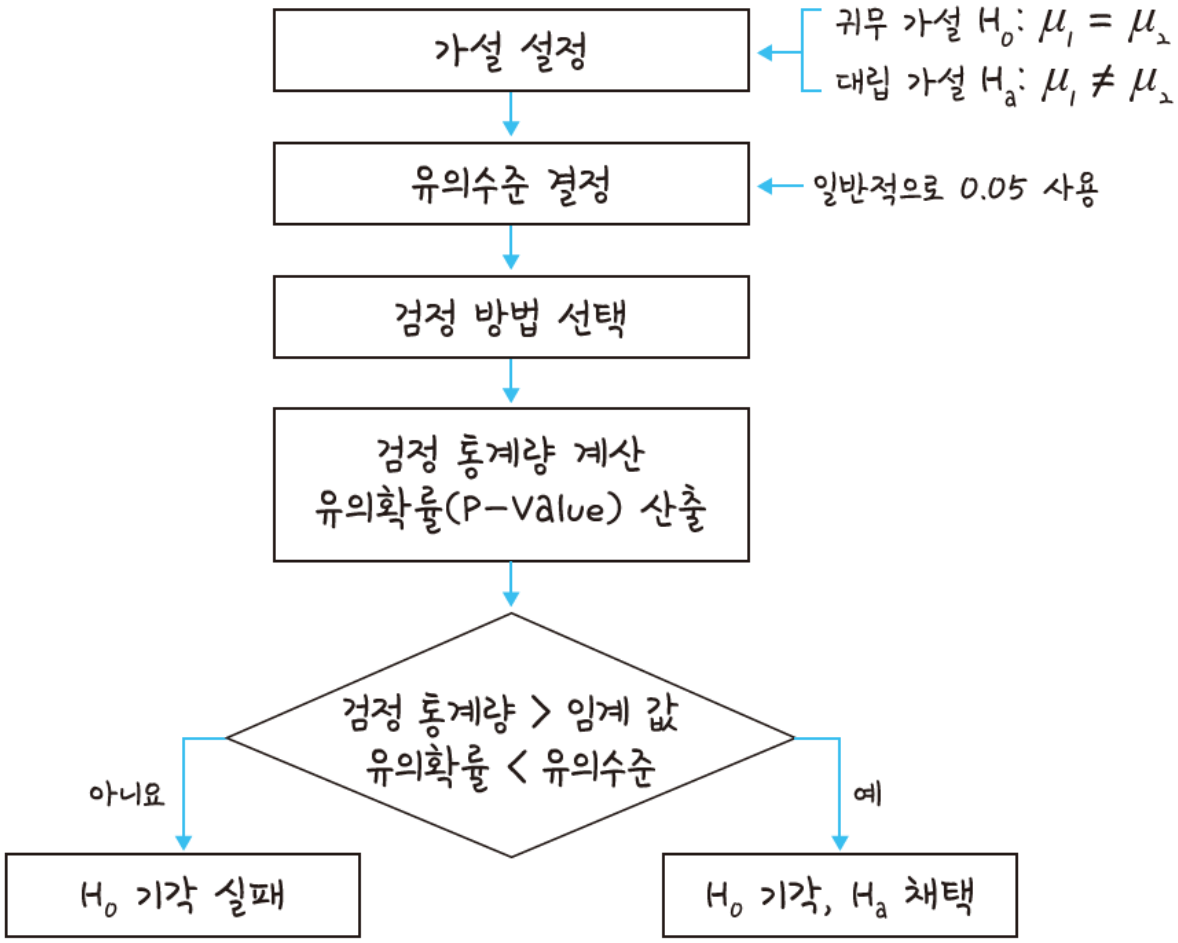
MATHEMATIC

❖ Statistical hypothesis test

Statistical hypothesis test procedure

- Statistical hypothesis test is one of the statistical estimation, the process of determining whether a hypothesis is reasonable using sample information based on the hypothesis that the actual observation value of the population
- This is called a hypothetical test for short
- Simply, the process of verifying whether a hypothesis is true or false after assuming that it is true or not
- If the hypothesis matches the observed value, the underlying hypothesis is not rejected, and if the hypothesis and observed value are inconsistent, the underlying hypothesis is rejected

Statistical test



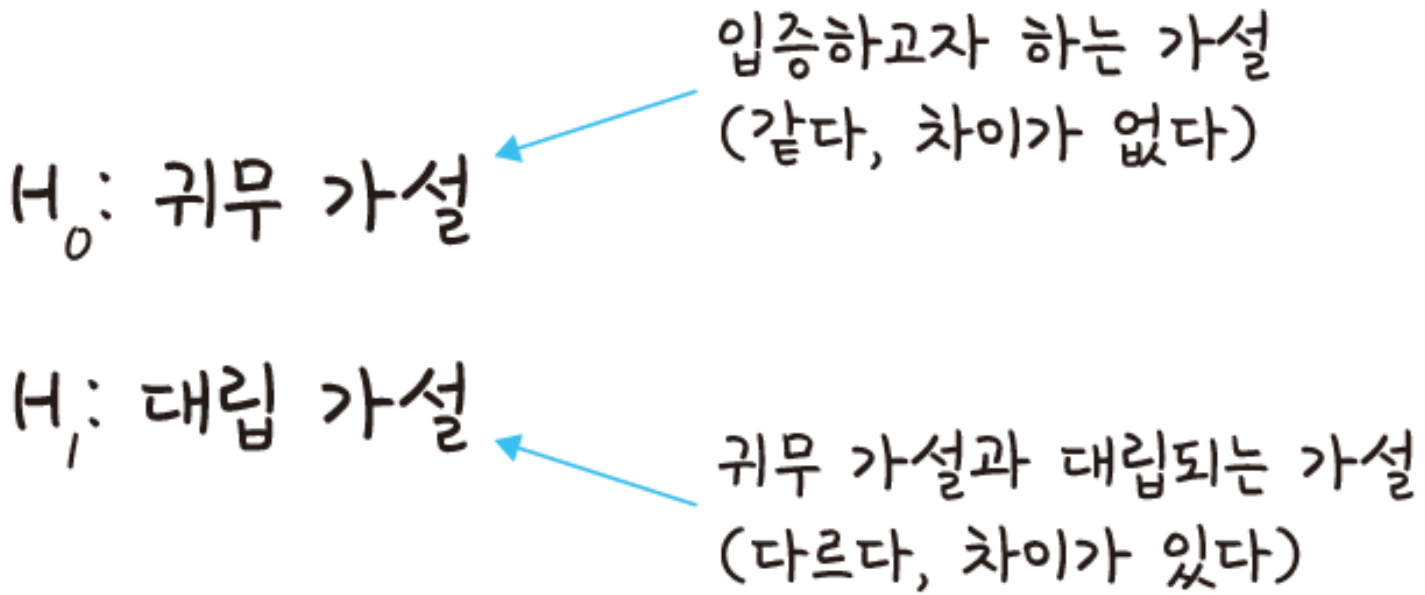
Statistical test

❖ Statistical hypothesis test

Hypothesis setting

- Hypothesis setting is the process of deriving potential answers to scientifically verify problems recognized based on observations of objects or data
- Hypothesis setting is very important because if you don't set a hypothesis, the model (algebra) is completed, and even if you complete the analysis, you can't determine if it's meaningful data
- By converting the research hypothesis "How to establish a hypothesis?" into a statistical hypothesis, we start testing the hypothesis
- Statistical hypotheses are presented as opposing hypotheses in opposition to the null hypothesis

Statistical test



Statistical test

MATHEMATIC

❖ Statistical hypothesis test

Null hypothesis

- The null hypothesis ($\mathbf{H}_0$) is a hypothesis that you want to prove, and you can think of it as a hypothesis that you want to prove
- The expression should be "no difference from", "no effect of", and "like“
- Due to the nature of the null hypothesis, you must choose the one that is more likely to occur naturally

Statistical test

❖ Statistical hypothesis test

Alternative hypothesis

- Alternative hypothesis ($\mathbf{H}_1$) is the opposite of the null hypothesis, and naturally the expression becomes "different", "different", "there is an effect", and "different“
- It means that it is less likely to occur naturally

Statistical test

❖ Statistical hypothesis test

Types of error

- Type 1 error is "an error that rejects the null hypothesis even though it is true," which is said to be effective even though it is not effective
- Type 2 error is 'an error that rejects the alternative hypothesis even though the alternative hypothesis is true', and it is said that it is actually effective, but it is not effective

검정 결과	실제	
	H_0(귀무 가설)이 참	H_0(귀무 가설)이 거짓
채택	참 확률 = $1-\alpha$	거짓(제2종 오류) 확률 = β(β위험)
기각	거짓(제1종 오류) 확률 = α(유의수준)	참 확률 = $1-\beta$

Statistical test

MATHEMATIC

❖ Statistical hypothesis test

- Let's look at example of type 1 and type 2 errors for determining whether cancer is diagnosed
  - Type 1 error: Decided it's cancer when it's not cancer
  - Type 2 error: It's cancer, but it's determined that it's not cancer

실제 상태 / 테스트 결과	암	암이 아님
'암'으로 판정	'암'을 '암'으로 판정	'암이 아님'인데 '암'이라고 판정(2종 오류)
'암이 아님'으로 판정	암인데, 알아내지 못함(1종 오류)	'암이 아님'을 '암이 아님'으로 판정

Statistical test

❖ Statistical hypothesis test

Determination of significance level

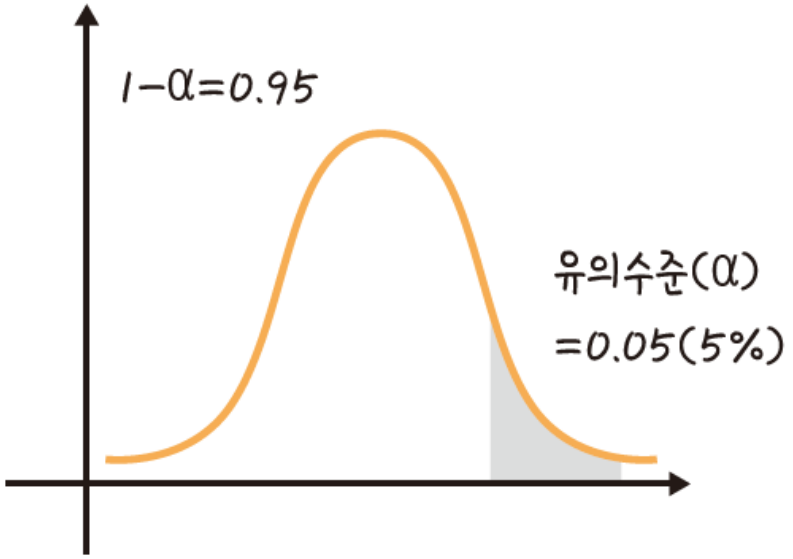
Significance level

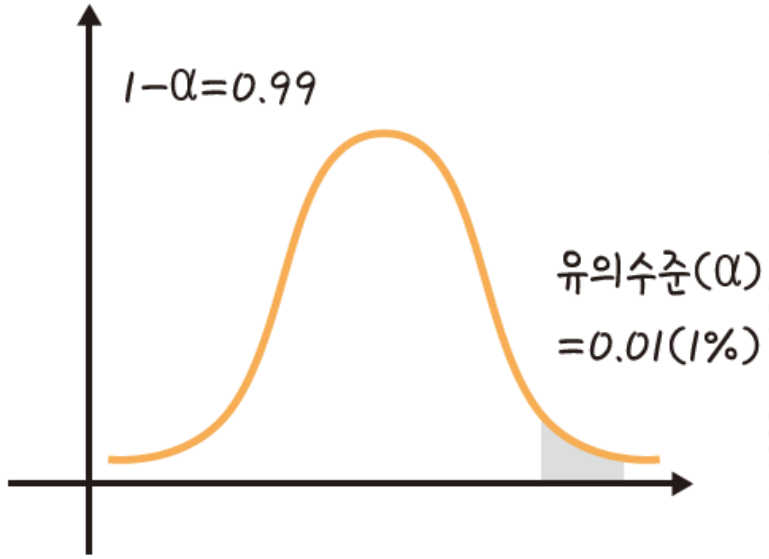
- In a hypothesis test, the probability of rejecting the null hypothesis and adopting the alternative hypothesis is called the significance level and is expressed as ***alpha* (α)**
- The significance level refers to the probability level that indicates how difficult the statistical value obtained from the sample is to obtain under the premise that the null hypothesis is correct
- If the significance level (α) is set to 0.05, the calculated significance probability (*p*-value) must be less than 0.05, the experimenter can reject the null hypothesis and adopt the alternative hypothesis

Statistical test

❖ Statistical hypothesis test

- The left side has an error tolerance of 5%, which is widely used in general social statistics, and the right side has an error tolerance of 1%, which is widely used in high-precision research fields

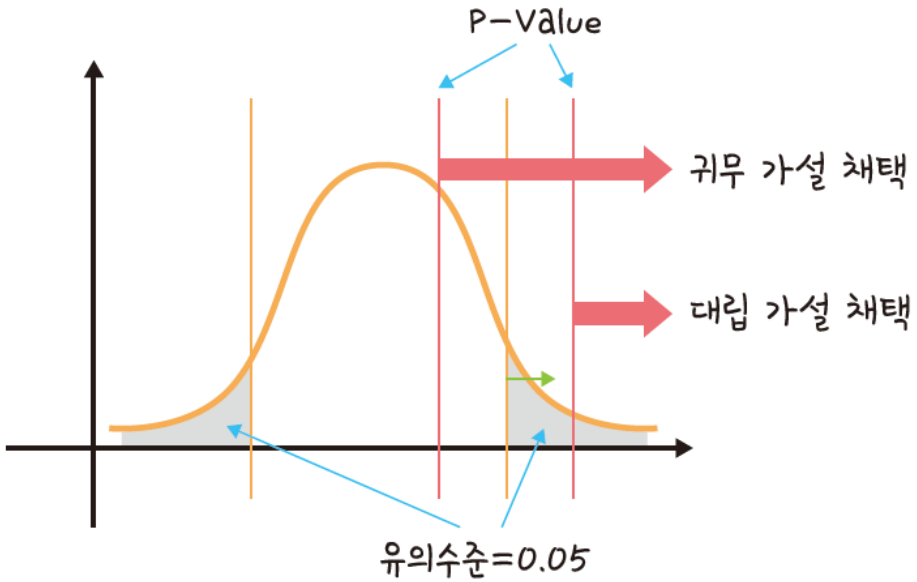




❖ Statistical hypothesis test

P-value

- *P*-value is called the significance probability, and means the minimum probability of rejecting the null hypothesis
- Based on the significance level, if the significance probability is higher than the significance level, the null hypothesis is adopted, and if it is low, the alternative hypothesis is adopted



Statistical test

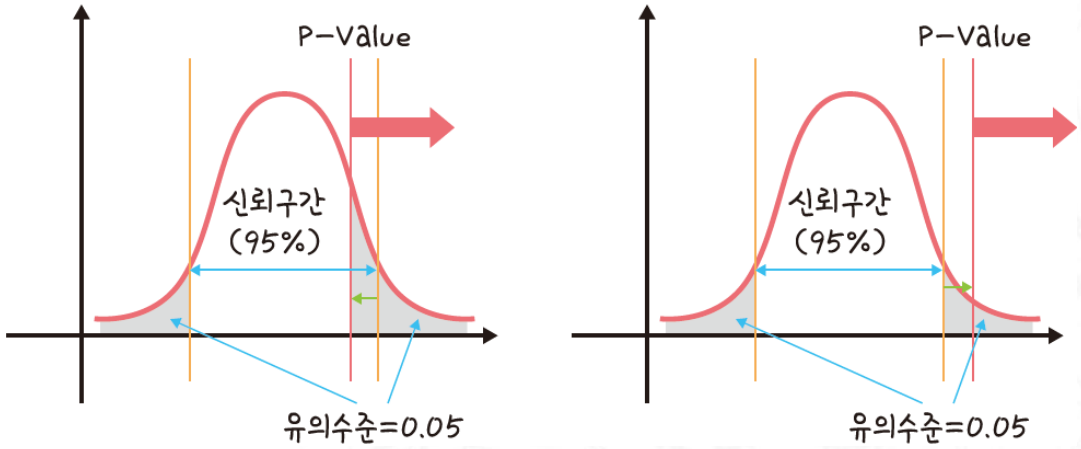
❖ Statistical hypothesis test

- Under the premise that the null hypothesis is correct, the probability of exceeding the current statistical value is a *p*-value
- If the *p*-value is too low, the probability of the hypothesis occurring is also low, so you reject the null hypothesis
- The standard is usually determined, but in general social statistics, it is based on 0.05 or 0.01
- This criterion is the level of significance learned earlier
  - Significance level: 0.05
    - *p*-value ≥ 0.05: Adopting the null hypothesis
    - *p*-value < 0.05: Adopting the alternative hypothesis

Statistical test

❖ Statistical hypothesis test

- As shown on the left, the larger the *p*-value, the more it is included in the confidence interval and the hypothesis is adopted
- On the other hand, the smaller the *p*-value, as on the right, the more out of the confidence interval and the hypothesis was rejected
- For reference, the value of 0.05, which is the criterion for adopting and rejecting p-values, means 5% (= 0.05), the remainder of the probability of being included in the confidence interval of 95% (= 0.95)



Statistical test

❖ Statistical hypothesis test

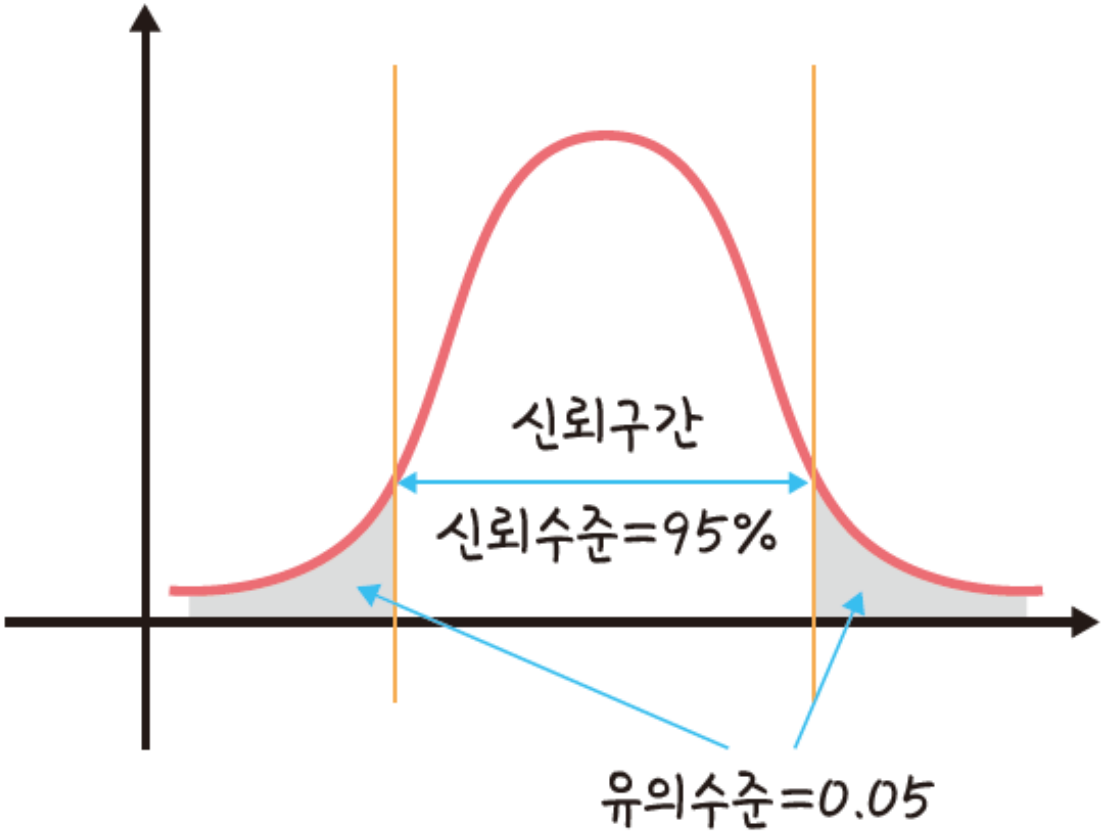
Confidence level and confidence interval

- In statistics, a sample is taken from a population and parameters (population mean, population standard deviation) are estimated using the statistics of the sample (sample mean, sample standard deviation)
- This is called ***point estimation***
- Since point estimation alone does not know how accurately the parameter is estimated, interval estimation is used to estimate the interval in which the population mean will exist, and the representative interval estimation is confidence interval
- The confidence interval is the interval in which the parameter is expected to be included, i.e., to show probabilistically what range the parameter is in
- Confidence intervals can be estimated in various sections such as 99% and 90%, but 95% confidence intervals are used a lot

Statistical test

❖ Statistical hypothesis test

- Confidence level is the rate at which the true value is in a certain range when many repetitions of the action of finding the true value are repeated



Statistical test

MATHEMATIC

❖ Statistical hypothesis test

Calculate test statistic (Sampling)

- Test statistics are random variables calculated from sample data and used for hypothesis testing
- Test statistics can be used to determine whether the null hypothesis is rejected

가설 검정	검정 통계량
z-검정	z-통계량
t-검정	t-통계량
분산 분석	F-통계량
카이제곱 검정	카이제곱 통계량

Statistical test

MATHEMATIC

❖ Statistical hypothesis test

- Use the z-test if you know the variance of the population, or the t-test if you don’t
- In realistic situations, most people don't know the population, and most often only know some of the "sample" extracted
- Most are verified by t-test
- The verification method for t-test is as follows
  - (1) Calculate the t-test statistic value for the sample (you get a value between -4 and 4)
  - (2) Show that the population and the mean are different as the values belong at both ends of the t-distribution graph
  - (3) The significance value is selected and tested using the specified p-value(%)

Statistical test

MATHEMATIC

❖ Statistical hypothesis test

- Difference between t-test and z-test

구분	t-검정	z-검정
사용 목적	두 집단 간 평균을 비교하는 검정 용도로 사용	모집단 평균의 차이를 검정하는 용도로 사용
언제 사용	모집단의 표준편차를 모를 때 사용	모집단의 표준편차를 알고 있을 때 사용
표본 크기	작음(30개 이하)	큼(30개 이상)

Statistical test

❖ Statistical hypothesis test

- The test statistic can be obtained by the following formula, which rejects or does not reject the null hypothesis depending on where this value is located in the probability distribution

$$\text{검정 통계량} = \frac{\text{표본평균} - \text{모평균}}{\text{표본표준편차}}$$

Statistical test

MATHEMATIC

❖ Statistical hypothesis test

In [16]:

```
import numpy as np
from scipy import stats

# 난수 발생을 위한 시드(seed) 1을 줍니다(코드를 실행할 때마다 똑같은 난수 생성)
np.random.seed(1)

# 평균 178, 표준편차 5로 임의의 높이 20개를 생성합니다
heights = [178 + np.random.normal(0, 5) for _ in range(20)]
```

Statistical test

MATHEMATIC

❖ Statistical hypothesis test

```
# t-검정 수행
tTestResult = stats.ttest_1samp(heights, 173)

# 결과 출력
print("The T-statistic is %.3f and the p-value is %.3f" % tTestResult)
```

```
The T-statistic is 3.435 and the p-value is 0.003
```

- Rejected the null hypothesis because the p-value is 0.003 when the regression region is set to $p < 0.05$
- This means that when the null hypothesis is true (when the average height of all students is 173cm), the probability of obtaining such a sample is 0.003, which means that the average height of students is not 173cm

Statistical test

MATHEMATIC

❖ Performance evaluation

- AI creates models based on data, and performance measurement is important because the accuracy of data classification (analysis) varies depending on the model performance
- Methods for measuring performance (1) the confusion matrix and (2) the ROC curve

Statistical test

MATHEMATIC

❖ Performance evaluation

Confusion matrix

- Let's find out four concepts that understand the confusion matrix

실제(condition) / 예측(prediction)	Positive	Negative
Positive	TP	FP
Negative	FN	TN

① TP (True Positive): What you predicted to be right

② TN (True Negative): What you predicted was wrong

③ FP (False Positive): What it predicted to be wrong

④ FN (False Negative): What you predicted was wrong

Statistical test

❖ Performance evaluation

- Using this concept, let's find out the performance evaluation method of confusion matrix
- To performance evaluation on a low-confusion matrix: Precision, Recall, and Accuracy
- Let's say we're diagnosing a patient's cancer, for example
- Precision, Recall, and Accuracy are as follows
  - Precision: The percentage of people who predict cancer patients who turn out to be real cancer patients, the formula is as follows

$$\text{정밀도(Precision)} = \frac{TP}{TP + FP}$$

Statistical test

❖ Performance evaluation

- Recall: The percentage of true cancer patients who are diagnosed as cancer patients

$$\text{재현율(Recall)} = \frac{TP}{TP + FN}$$

- Accuracy: As a percentage of all cancer patients who identify as cancer patients

$$\text{정확도(Accuracy)} = \frac{TP + TN}{TP + FP + TN + FN}$$

Statistical test

❖ Performance evaluation

- For reference, if precision and recall are viewed separately, it may be difficult to determine due to a trade-off, and in this case, the performance can be evaluated using the harmonic average of the two
- This is the F1 score (F1 Score), which is the following formula

$$F_1 = 2 \times \frac{\text{정밀도} \times \text{재현율}}{\text{정밀도} + \text{재현율}}$$

Statistical test

❖ Performance evaluation

- Python provides accuracy, precision, and recall as follows

In [17]:

```
# 혼동행렬을 위한 sklearn 라이브러리를 호출합니다
import numpy as np
import sklearn.metrics as metrics

y = np.array([1, 1, 1, 1, 0, 0]) # 0은 정상, 1은 암환자
p = np.array([1, 1, 0, 0, 0, 0]) # 예측 값

# sklearn(sklearn.metrics)을 이용하여 정확도, 정밀도, 재현율, F1 스코어
를 계산합니다
# accuracy_score() 함수로 정확도를 계산합니다
print('accuracy', metrics.accuracy_score(y,p))
```

Statistical test

MATHEMATIC

❖ Performance evaluation

```
# precision_score() 함수로 정밀도를 계산합니다
print('precision', metrics.precision_score(y,p))
print('recall', metrics.recall_score(y,p)) # recall_score() 함수로
재현율을 계산합니다
print('f1', metrics.f1_score(y,p)) # f1_score() 함수로 F1 스코어를
계산합니다

# 정확도, 정밀도, 재현율, F1 스코어를 한 번에 출력합니다
print(metrics.classification_report(y,p))
print(metrics.confusion_matrix(y,p))
```

Statistical test

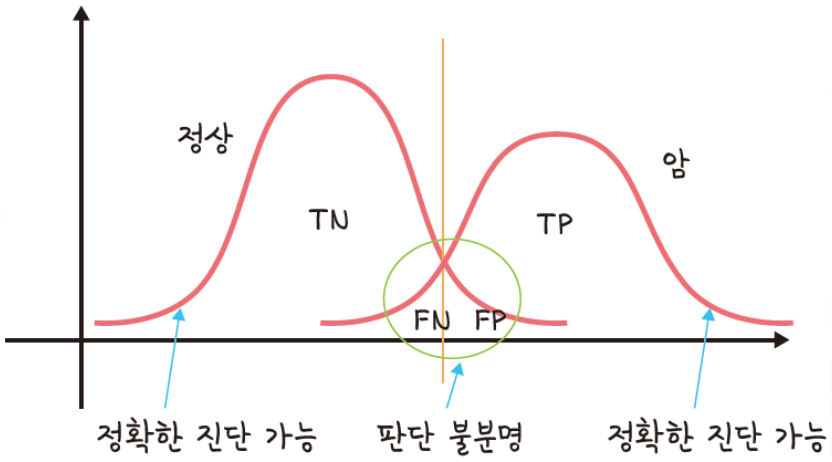
❖ Performance evaluation

```
accuracy 0.6666666666666666
precision 1.0
recall 0.5
f1 0.6666666666666666
              precision    recall  f1-score   support

           0       0.50      1.00      0.67         2
           1       1.00      0.50      0.67         4

    accuracy                           0.67         6
   macro avg       0.75      0.75      0.67         6
weighted avg       0.83      0.67      0.67         6

[[2 0]
 [2 2]]
```

❖ Performance evaluation

ROC curve

- The ROC curve is used to compensate for the shortcomings of the confusion matrix
- When the distribution of the two classes is different, accurate diagnosis is possible at both ends, but the accuracy is poor because the judgment is unclear in the middle

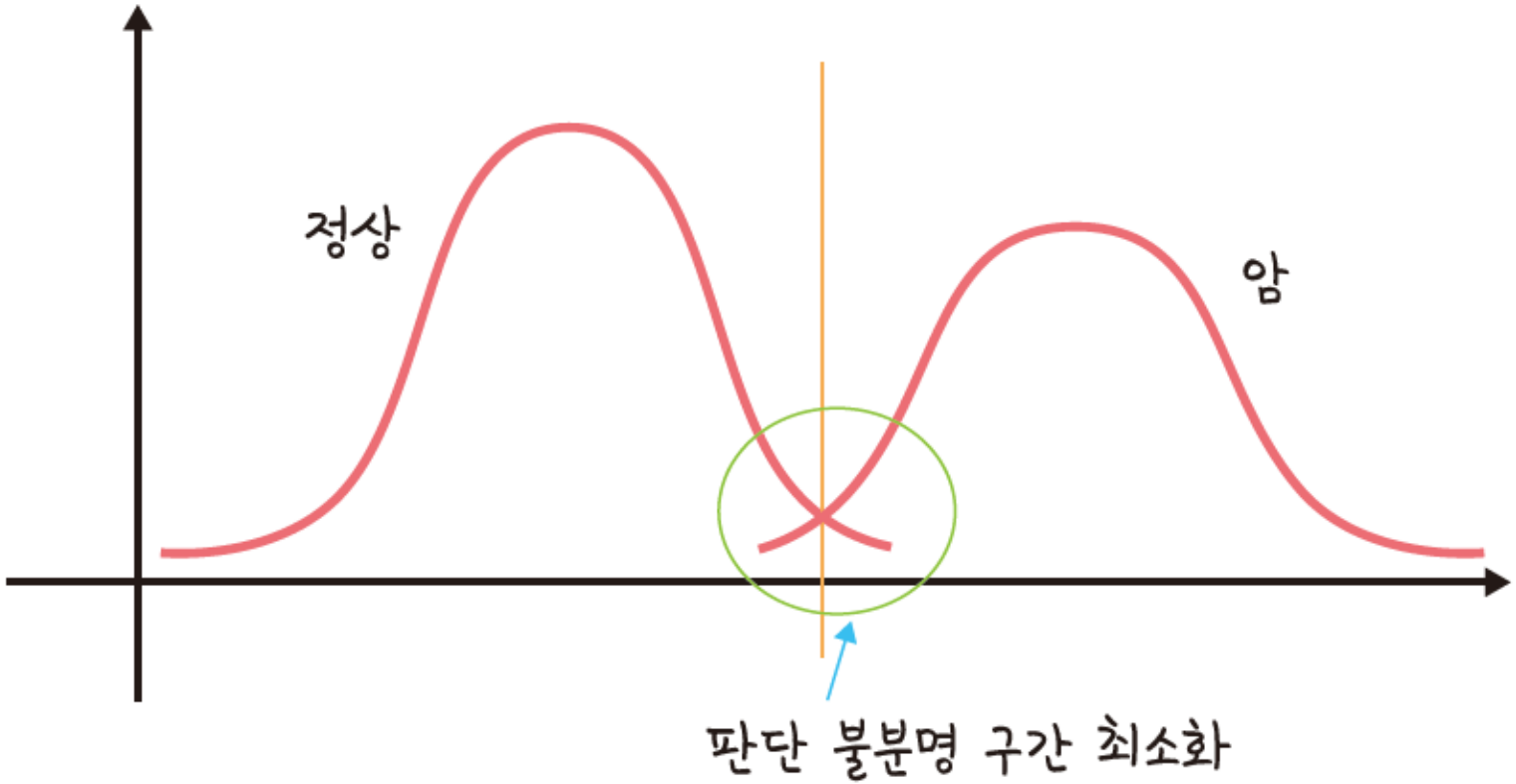


Statistical test

MATHEMATIC

❖ Performance evaluation

- The ROC curve minimizes the unclear section of the judgment



Statistical test

MATHEMATIC

❖ Performance evaluation

- The y-axis of the ROC curve represents the true positive rate (TPR), and the x-axis represents the false positive rate (FPR)
- The area below is called AUC (Area Under the Curve), and the larger the AUC area, the better the curve
  - Sensitivity : a prediction of what is right
  - Specificity : the prediction that something is wrong
  - AUC : the area below the graph, digitizing the calculation results for simple comparison of performance

Statistical test