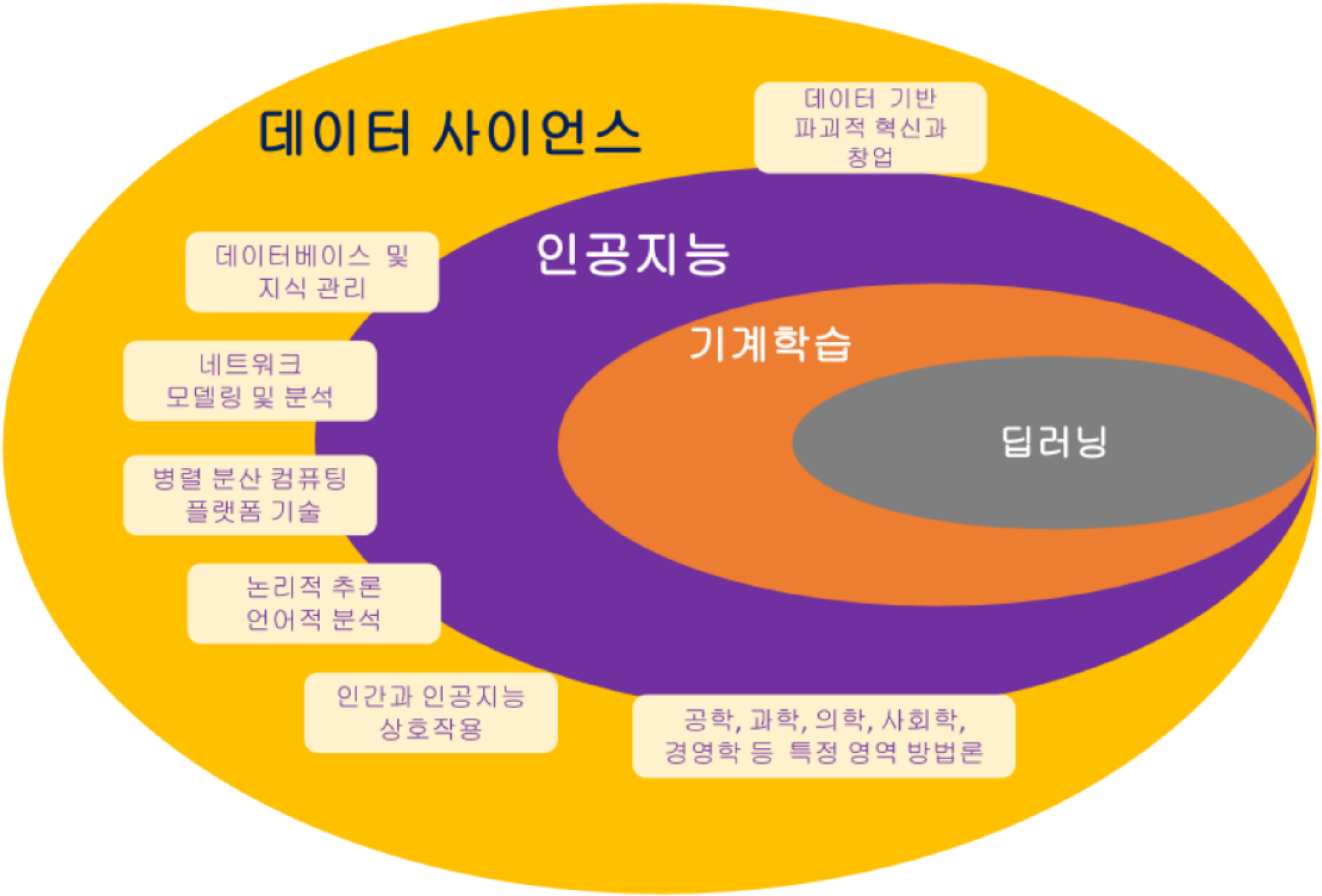
기초 수학 통계

Contents

데이터 사이언스(Data Science)란?



빅데이타(Big Data)란?

인공지능 VS 머신러닝 VS 딥러닝

출처 : https://hyeonjiwon.github.io/assets/img/21-1.png

머신러닝, 딥러닝 분야에서 수학 지식이 필요할까요?

LINEAR ALGEBRA

DERIVATIVE

STATISTICS

PROBABILITY

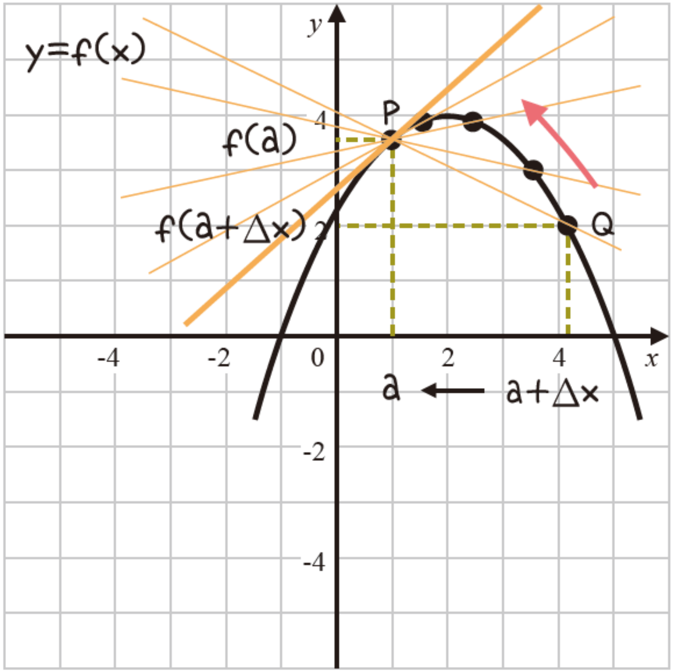
머신러닝, 딥러닝 분야에서 수학 지식이 필요할까요?

$$\mathbf{A} = \left[a_{jk}\right] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} : \ m \times n$$

머신러닝, 딥러닝 분야에서 수학 지식이 필요할까요?

Numerical derivative

$$f'(x) = \frac{df(x)}{dx} \qquad \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$



머신러닝, 딥러닝 분야에서 수학 지식이 필요할까요?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

추천 사이트

- 칸아카데미(https://ko.khanacademy.org)
- 한양대학교 OCW (http://ocw.hanyang.ac.kr)
- 확률 및 통계

  (http://www.kocw.net/home/search/kemView.do?kemId=1056974)
- 선형대수학(3Blue1Brown)

  (https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab)

쥬피터 노트북

- Jupyter notebook은 대화형 파이썬 인터프리터(Interpreter) 도구로 아나콘다 설치시 함께 설치된다.
- 구글 크롬과 같이 웹 브라우저 환경에서 파이썬 코드를 작성 및 실행할 수 있는 도구이다.
- 서버에 설치하거나 로컬 환경에서 포트를 개방한 후 해당 url에 접속하여 사용이 가능하다.

쥬피터 노트북 실행

바로가기 아이콘

1. 노트북 실행 파일 위치 찾기

탐색기에서 아래 위치의 파일을 탐색한다.

C:\ProgramData\Anaconda3\Scripts\jupyter-notebook.exe

마우스 우측버튼 클릭후 [보내기]-[바탕화면에 바로가기만들기]

2. 1번에서 생성된 바로가기아이콘을 마우스 우측 버튼
   [속성]

3. 속성 대화상자에서
   [시작위치] 부분을 작업폴더 위치로 지정
   예) C:\pyclassD

쥬피터 노트북 실행

아나콘다 프롬프트 이용

1) 아나콘다 프롬프트 실행

- [Anaconda3]-[Anaconda3 Prompt ...]를 마우스 우측 버튼
- [자세히]-[관리자 권한으로 실행]
- [Anaconda3 Prompt ...] 터미널 모드 확인

2) 터미널 모드 명령 실행 : 디렉토리 위치 변경

cd 이동할작업폴더경로

예) cd c:\pythonClass

3) 쥬피터노트북 실행 명령

jupyter notebook

쥬피터노트북 명령어

쥬피터 노트북의 파일 확장자는? *.ipynb

쥬피터 노트북의 명령셀 모드

- Code : 일반적인 파이썬 명령 모드
- MarkDown : 주석모드

명령셀 선택 상태에서

m : 마크다운 셀 모드

y : 명령 셀 모드

쥬피터노트북 명령어

- **많이 쓰는 단축키**
- Ctrl+Enter : 셀 실행
- Shift+Enter : 셀 실행 후 아래 셀 선택
- Alt+Enter : 셀 실행 후 아래에 셀삽입
- Ctrl-s : 노트북 저장
- Ctrl-z : undo
- Ctrl-y : redo
- a : 위에 cell 추가
- b : 아래 cell 추가
- x : cell 잘라내기
- v : 아래에 cell 붙여넣기

LaTex 문법

위키백과:LaTeX 문법
https://ko.wikipedia.org/wiki/%EC%9C%84%ED%82%A4%EB%B0%B1%EA%B3%BC:TeX_%EB%AC%B8%EB%B2%95

```
$$ y=x^3+2x^2+x+3 $$
```

$$y = x^3 + 2x^2 + x + 3$$

```
$$ \frac{2}{4} or {2 \over 4} $$
```

$$\frac{2}{4} or {2 \over 4}$$

LaTex 문법

수학변수와 그리스 문자

영어	한글	입력방법	문자
alpha	알파	`$\alpha$, $$\alpha$$`	α
beta	베타	`$\beta$, $$\beta$$`	β
gamma	감마	`$\gamma$, $$\gamma$$`	γ
delta	델타	`$\delta$, $$\delta$$`	δ
epsilon	엡실론	`$\epsilon$, $$\epsilon$$`	ϵ
zeta	제타	`$\zeta$, $$\zeta$$`	ζ
eta	에타	`$\eta$, $$\eata$$`	η
theta	쎄타	`$\theta$, $$\theta$$`	θ
kappa	카파	`$\kappa$, $$\kappa$$`	κ
lamdba	람다	`$\lambda$, $$\lambda$$`	λ
mu	뮤	`$\mu$, $$\mu$$`	μ

LaTex 문법

수학변수와 그리스 문자

nu	누	ν, $$\nu$$	ν
xi	크싸이	ξ, $$\xi$$	ξ
pi	파이	π, $$\pi$$	π
rho	로	ρ, $$\rho$$	ρ
sigma	시그마	σ, $$\sigma$$	σ
tau	타우	τ, $$\tau$$	τ
phi	파이	ϕ, $$\phi$$	ϕ
chi	카이	χ, $$\chi$$	χ
psi	프사이	ψ, $$\psi$$	ψ
omega	오메가	ω, $$\omega$$	ω

LaTex 문법

```
$ y $  실제값  $\hat{y}$  예측값  $\bar{y}$ 평균값
```

y 실제값 $\hat{y}$ 예측값 $\bar{y}$ 평균값

$$\sum_{k=1}^{N} k^2$$

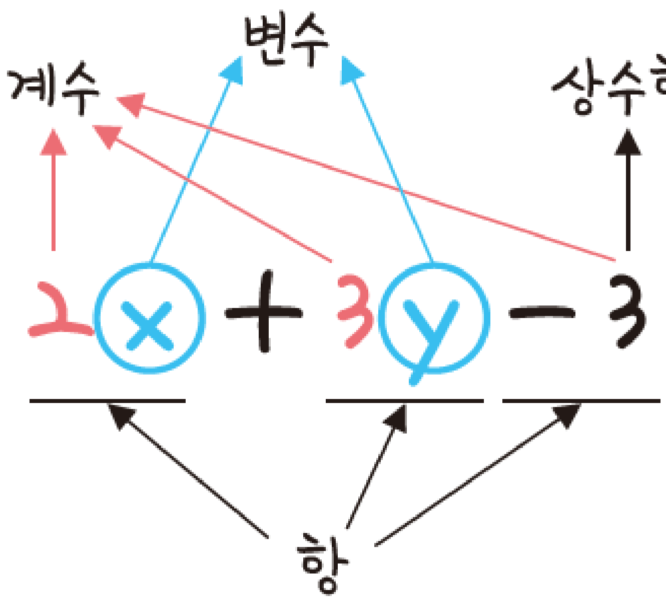
```
$$ ₩sum_{k=1}^N k^2 $$
```

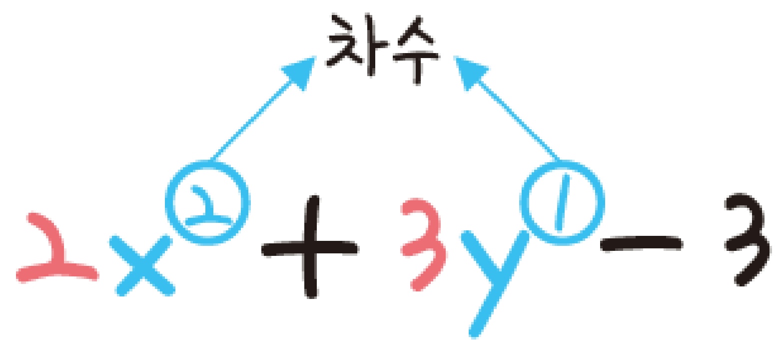
방정식

방정식(equation) 이란?

$$x + 2 = 6$$

방정식





방정식

이항

(1) $x + 1 = 3$

(2) $x + 1 - 3 = 0$

이항하면 기호가 바뀜

(3) $x - 2 = 0$

(일차방정식) = 0 형태이므로 일차방정식임

SymPy

Sympy란 기호 수학(symbolic math)을 위한 파이썬 라이브러리

```
> pip install sympy   또는   conda install sympy
```

https://docs.sympy.org/latest/index.html

방정식 구현

```python
from sympy import Symbol, solve
```

```python
x = Symbol('x')
x
```

x

```python
equation = 2*x - 6
equation
```

$2x - 6$

```python
solve(equation)
```

[3]

방정식 연습문제

다음 방정식의 해를 구하세요.

(1) $4 = k - 2$

(2) $10 = 2k$

(3) $\dfrac{k}{2} = 8$

연립 방정식

$$\begin{cases} 3x + y = 2 & \text{------ (1)} \\ x - 2y = 3 & \text{------ (2)} \end{cases}$$

$$\begin{array}{r} 3x + \quad y = 2 \\ -\big)\ \boxed{3x - 6y = 9} \\ \hline 7y = -7 \end{array}$$

×3 취함

연립 방정식 구현

```
from sympy import Symbol, solve
```

```
x=Symbol('x')
y=Symbol('y')
x, y
```

```
(x, y)
```

```
equation1 = 3 * x + y - 2
equation2 = x - 2 * y - 3
```

```
solve((equation1, equation2), dict=False)
```

```
{x: 1, y: -1}
```

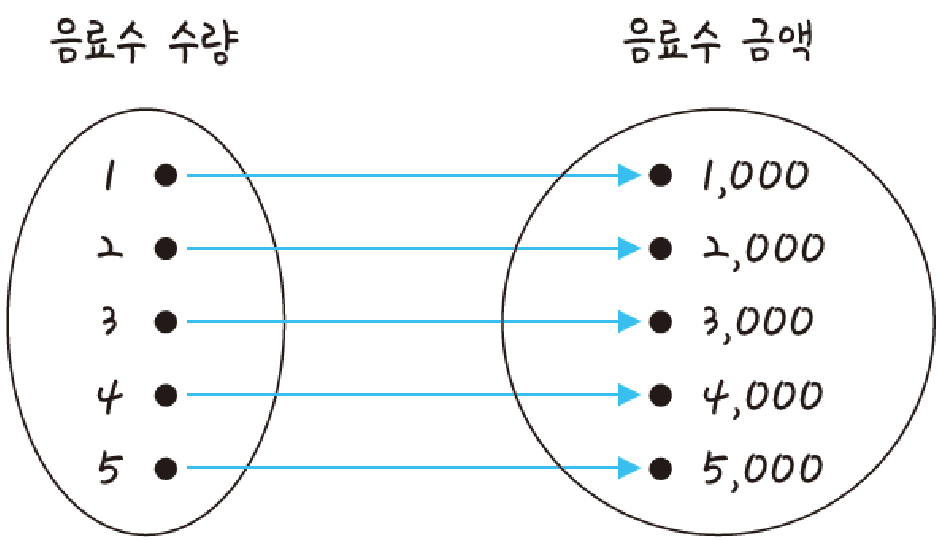
함수

함수란 첫 번째 집합의 임의의 한 원소를 두 번째 집합의 원소 하나에 대응시키는 관계

표 2-1 음료수 수량과 금액

음료수 수량	1	2	3	4	5
음료수 금액	1,000	2,000	3,000	4,000	5,000

그림 2-1 음료수 수량과 금액의 대응 관계



$$Y = 1000X$$

$$Y = f(X)$$

함수

➢ 일차함수와 그래프

일차함수 그래프

● 일차함수 y = ax 그래프

표 2-3 a가 양수일 때 변화

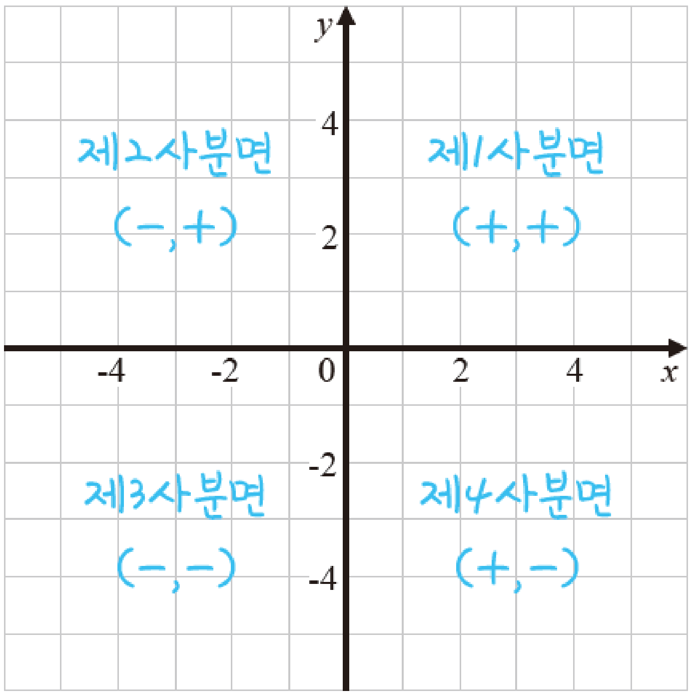
a 값	1	2	3
y 값	1x	2x	3x

표 2-4 a가 음수일 때 변화

a 값	–1	–2	–3
y 값	–1x	–2x	–3x

함수

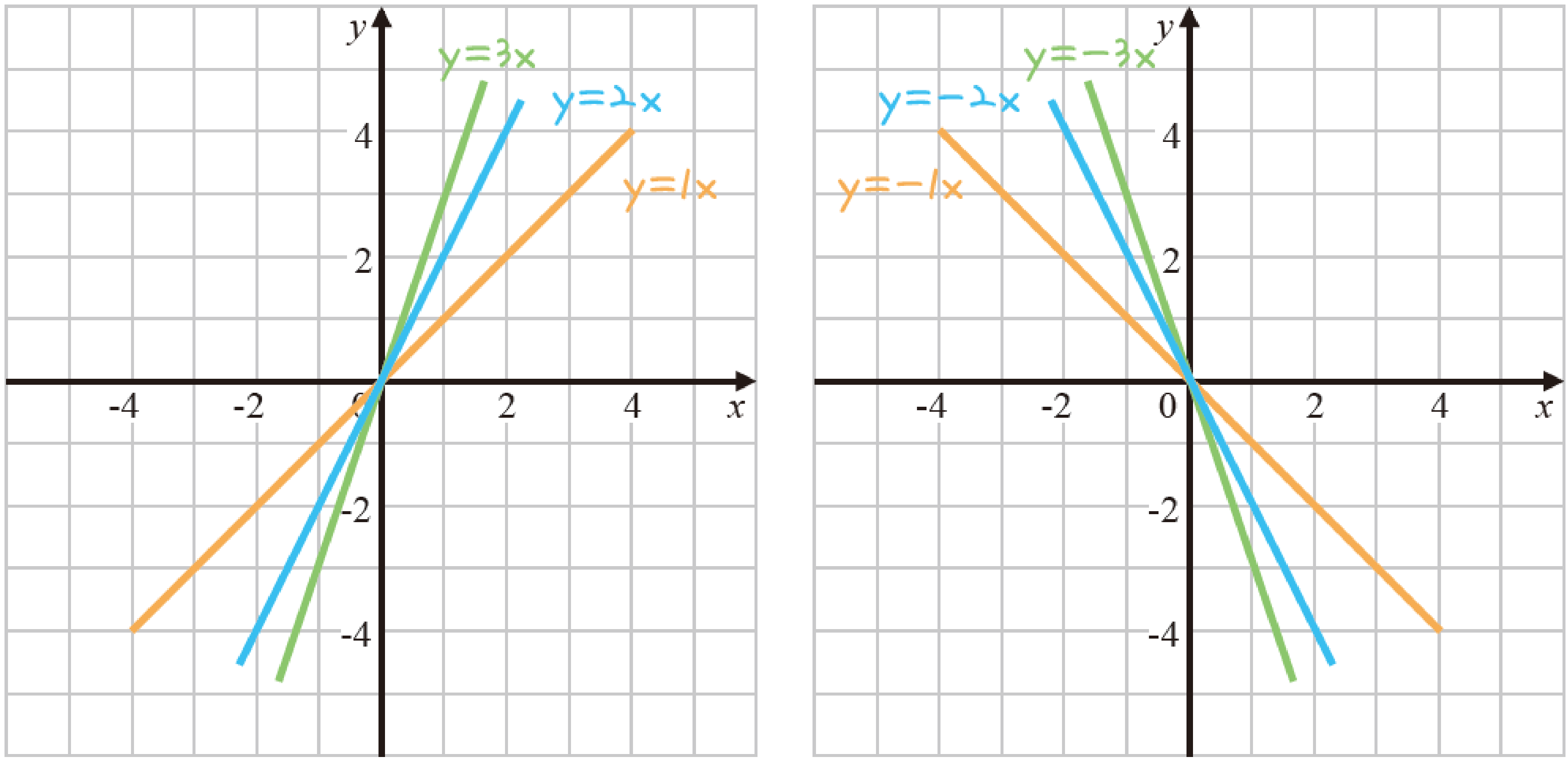
➢ 좌표 평면과 사분면



- 제1사분면: $x > 0$, $y > 0$을 만족하는 영역
- 제2사분면: $x < 0$, $y > 0$을 만족하는 영역
- 제3사분면: $x < 0$, $y < 0$을 만족하는 영역
- 제4사분면: $x > 0$, $y < 0$을 만족하는 영역

함수

그림 2-2 일차함수 y = ax 그래프



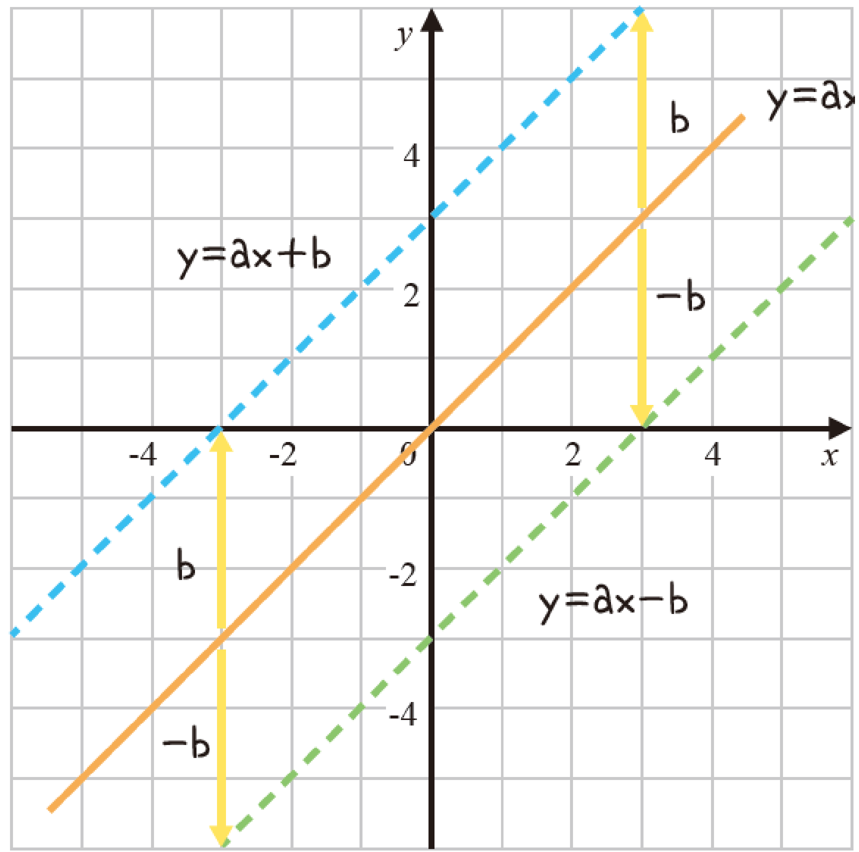
① y=ax 그래프(a가 양수)

② y=ax 그래프(a가 음수)

함수

➢ 일차함수와 그래프

$$y = \underset{\text{직선의 기울기}}{a}x + \underset{y\text{ 절편}}{b}$$

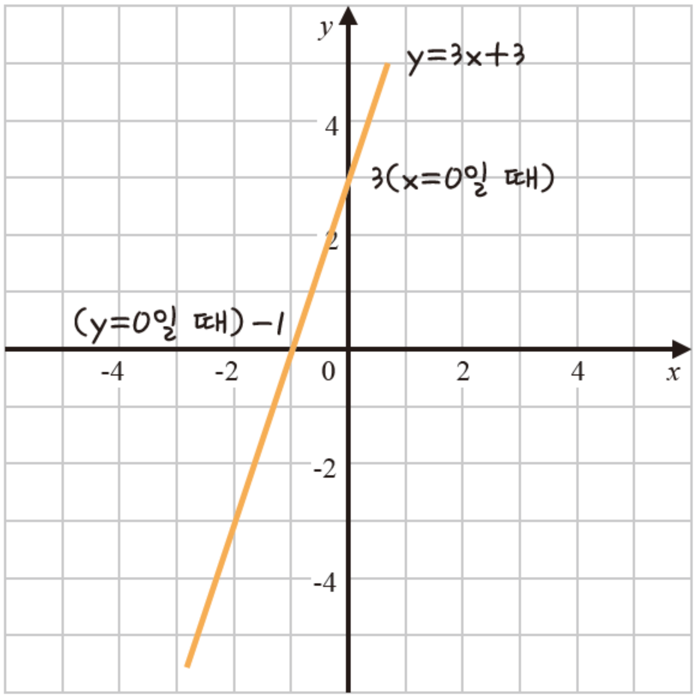


함수

➢ x 절편과 y 절편

- 절편이란?
  좌표 평면 위의 직선이 x축과 만나는 점(x 좌표)과 y축과 만나는 점(y 좌표)

그림 3-6 x 절편, y 절편

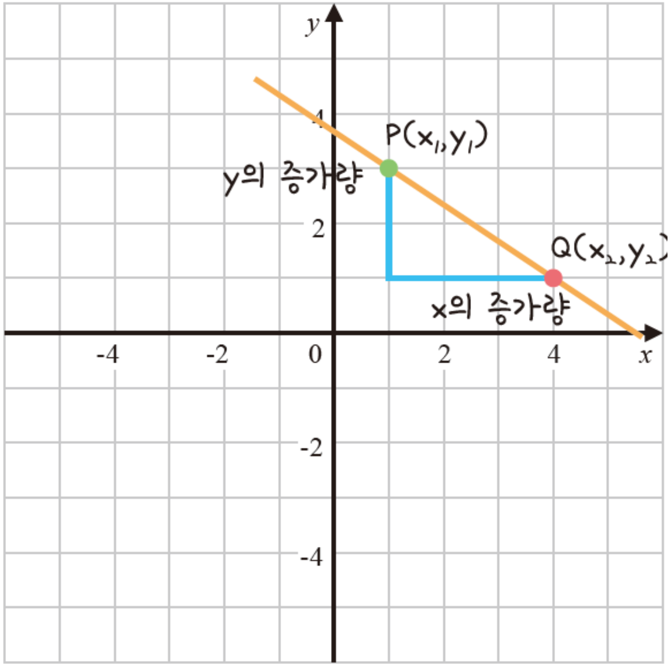


함수

➢ 기울기

$$기울기 = \frac{y의\ 증가량}{x의\ 증가량} = \frac{y_2 - y_1}{x_2 - x_1}$$

그림 3-7 기울기



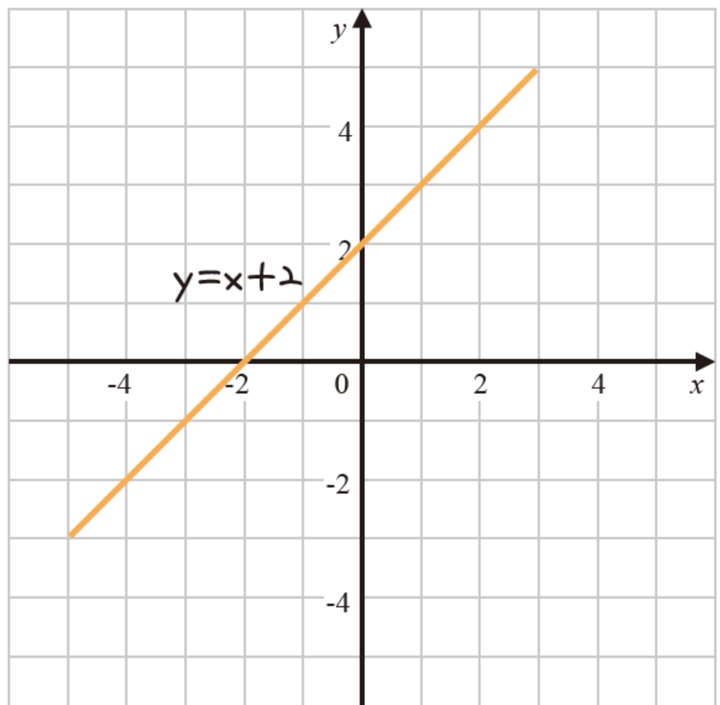
함수

➢ 기울기

(1) x축과 만나는 점의 좌표와 y축과 만나는 점의 좌표 => (-2, 0), (0, 2)

(2) 기울기 : $\frac{y_2 - y_1}{x_2 - x_1}$ 공식을 적 $\frac{2-0}{0-(-2)} = 1$ 이 됨

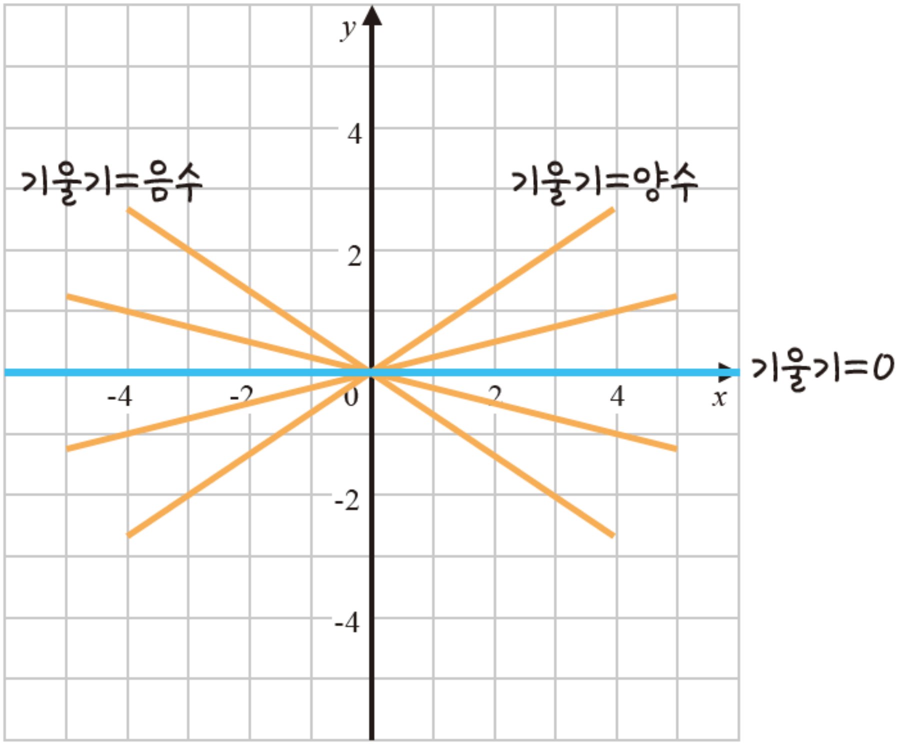
그림 3-8 y = x + 2 그래프



함수

➢ 양의 기울기와 음의 기울기

그림 3-9 양의 기울기와 음의 기울기



함수

➢ 이차함수와 그래프

$$y = f(x) = ax^2 + bx + c \qquad \text{수식 2.5}$$

그림 2-6 수식의 해와 절편

$$y = \underbrace{ax^2 + bx + \overset{\text{y 절편}}{\downarrow}\!\!\! c}_{\text{이 수식의 해, x 절편}}$$

함수

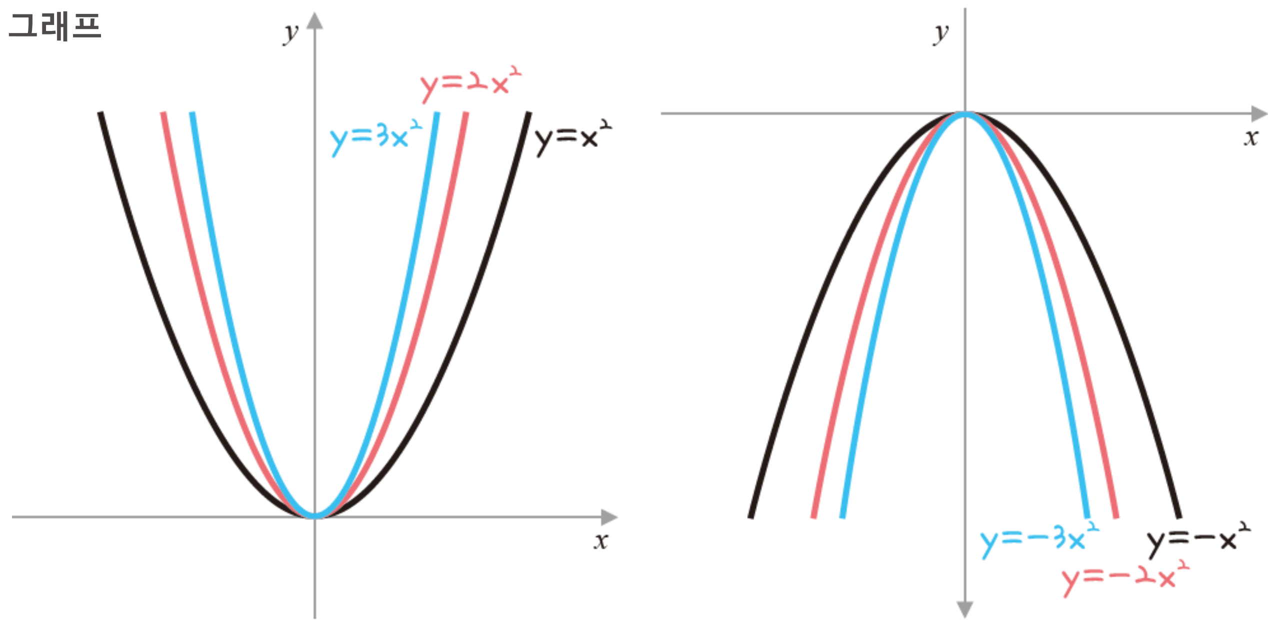
➢ 이차함수와 그래프

이차함수 y = ax**2 그래프

표 2-5 a가 양수일 때 변화

a 값	1	2	3
y 값	$1x^2$	$2x^2$	$3x^2$

그림 2-7 이차함수 y = ax2 그래프

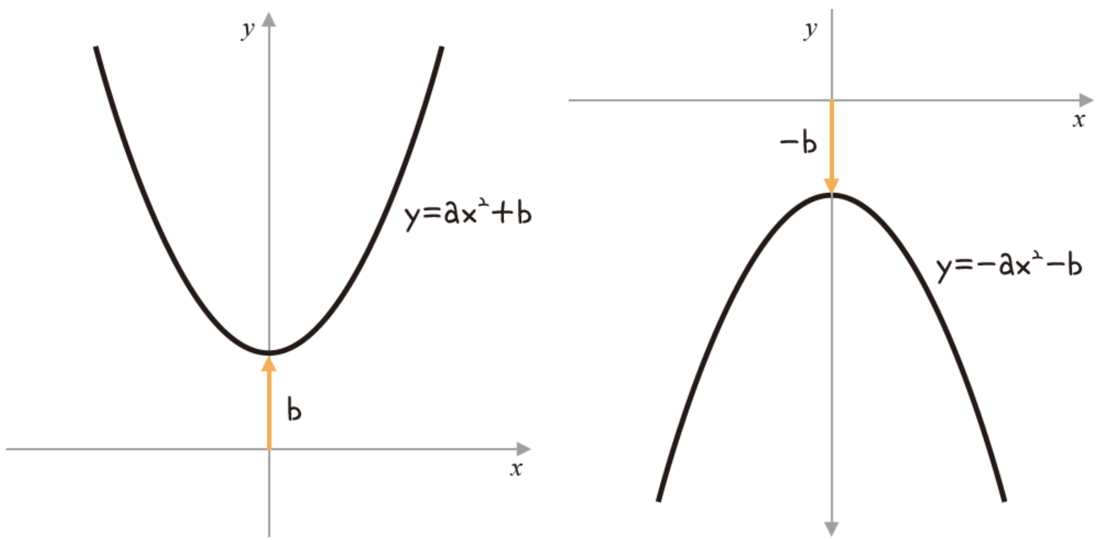


① $y=ax^2$ $(a>0)$ ② $y=ax^2$ $(a<0)$

함수

➢ 이차함수와 그래프

이차함수 y = ax**2 + b 그래프



① $y=ax^2+b$ ② $y=-ax^2-b$

1차 함수 파이썬 구현

```python
x = Symbol('x')
x
```

```python
fx = 3*x - 2
fx
```

$3x - 2$

```python
print(f' x = 1 일때 함수값은? {fx.subs(x, 1)}')
print(f' x = 5 일때 함수값은? {fx.subs(x, 5)}')
print(f' x = 10 일때 함수값은? {fx.subs(x, 10)}')
```

```
x = 1 일때 함수값은? 1
x = 5 일때 함수값은? 13
x = 10 일때 함수값은? 28
```

2차 함수 파이썬 구현

```python
x = Symbol('x')
x
```

x

```python
fx = x**2 + 3*x + 5
fx
```

$x^2 + 3x + 5$

```python
print(f' x = 1 일때 함수값은? {fx.subs(x, 1)}')
print(f' x = 5 일때 함수값은? {fx.subs(x, 5)}')
print(f' x = 10 일때 함수값은? {fx.subs(x, 10)}')
```

```
x = 1 일때 함수값은? 9
x = 5 일때 함수값은? 45
x = 10 일때 함수값은? 135
```

2차 함수 파이썬 구현

넘파이 데이터 준비하기

```
x = np.arange(-4, 5)
x
```

```
array([-4, -3, -2, -1,  0,  1,  2,  3,  4])
```

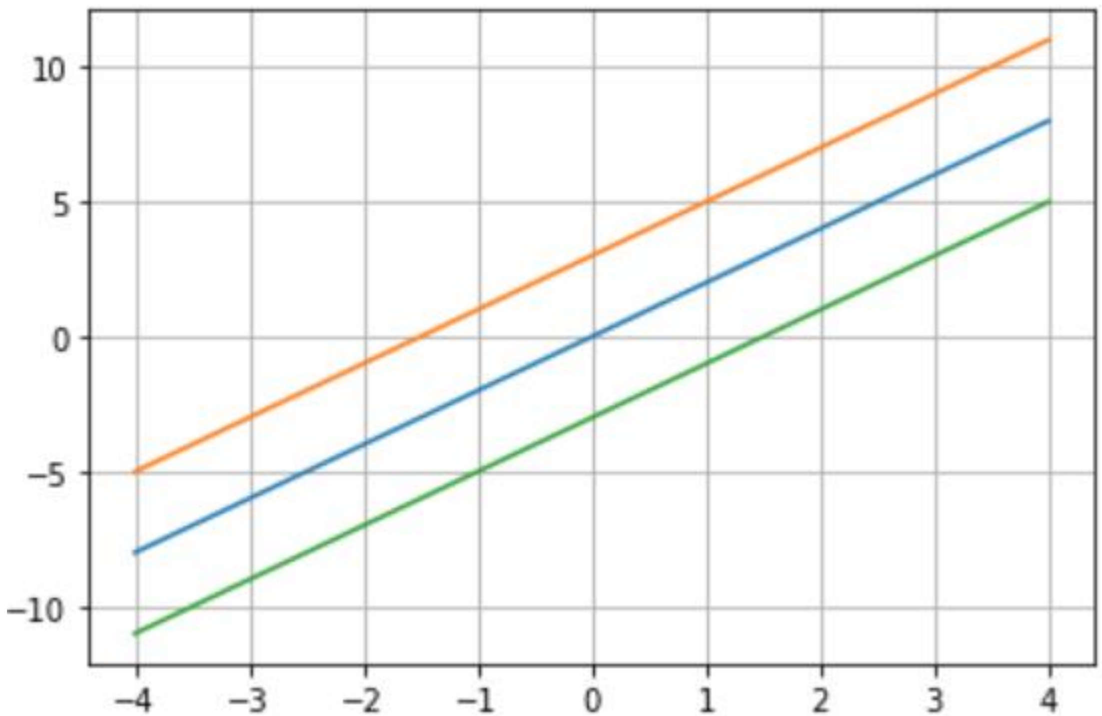
```
y = 2*x
y
```

```
array([-8, -6, -4, -2,  0,  2,  4,  6,  8])
```

1차 함수 그래프

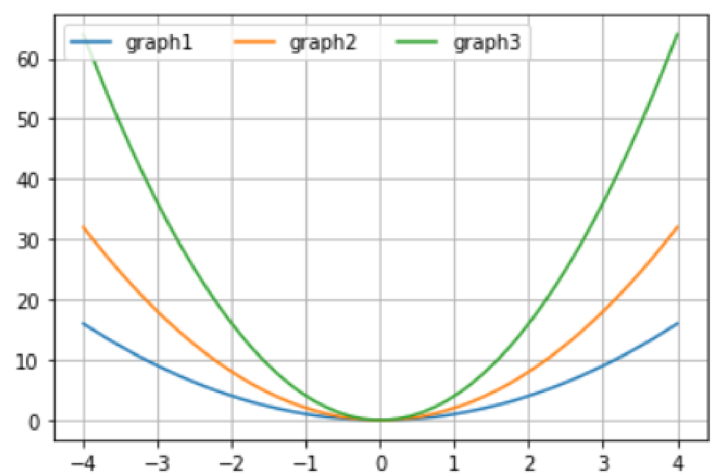
```python
import matplotlib.pyplot as plt
import numpy as np
```

```python
x = np.arange(-4, 5)
y = 2*x
plt.plot(x, y)
y = 2*x + 3
plt.plot(x, y)
y = 2*x - 3
plt.plot(x, y)
plt.grid()
plt.show()
```



2차 함수 그래프

```python
x = np.linspace(-4, 4, 50)
y = x**2
plt.plot(x, y, label='graph1')
y = 2*x**2
plt.plot(x, y, label='graph2')
y = 4*x**2
plt.plot(x, y, label='graph3')
plt.grid()
plt.legend(loc='upper left', ncol=3)
# plt.legend(loc='lower center')
plt.show()
```



미분

미분이란?

미분은 한 점에서의 기울기를 의미한다

f '(a) = a라는 점에서의 기울기
= a라는 점에서 접선의 기울기
= a라는 점에서의 미분 값
= a라는 점에서의 미분계수

$$\text{기울기} = \frac{y\text{의 증가량}}{x\text{의 증가량}} = \frac{f(a+h)-f(a)}{(a+h)-a}$$

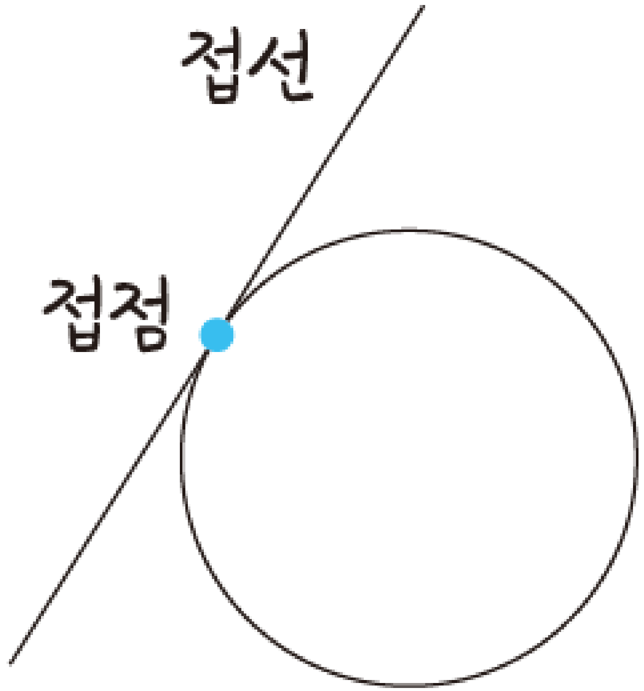
수식 7.1

$$\lim_{x\text{ 증가량}\to 0}\frac{y\text{의 증가량}}{x\text{의 증가량}} = \lim_{h\to 0}\frac{f(a+h)-f(a)}{(a+h)-a}$$

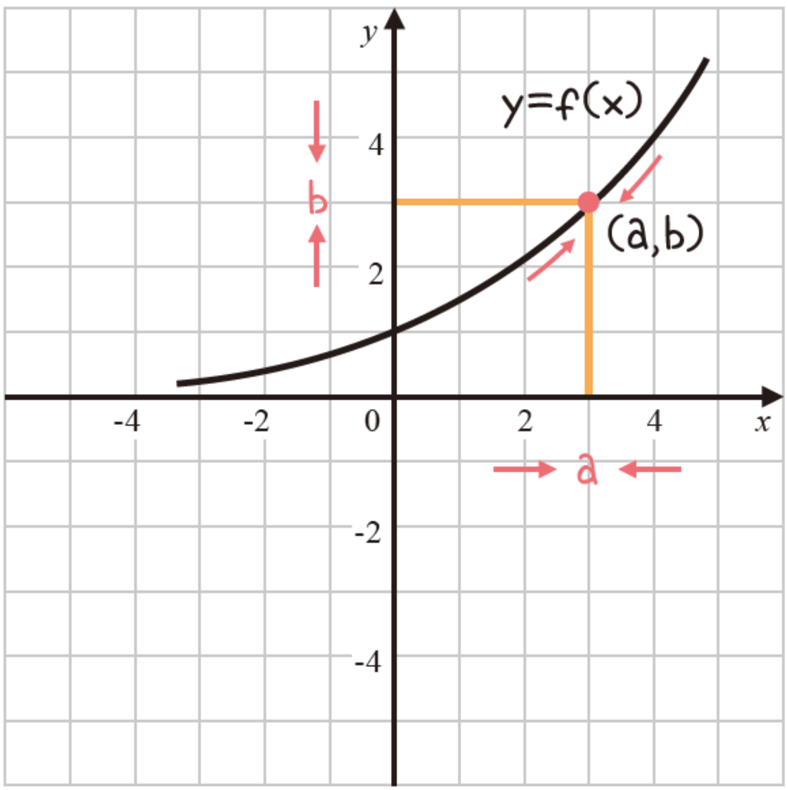
수식 7.2

미분

➤ **접선 :** 곡선 위의 한 점에서 곡선에 접하는 직선



▼ 그림 7-5 함수의 수렴



미분

➢ 평균변화율

- x가 변하는 양에 대해 y가 얼마나 변하는지를 구하는 변화율로 **평균변화율**과 **순간변화율**이 있다
- 순간변화율 : 찰나의 순간에 대한 변화율을 구하는 것으로 미분이라고도 하며, 그 찰나의 변화율을 순간변화율 혹은 **미분계수**라고도 한다.
- 수학에서 증가량을 표현할 때는 Δx 기호를 사용함
- 예를 들어 x가 1에서 5까지 변했다고 가정했을 때, x는 4만큼 증가했다면 이를 기호로 $\Delta x = 4$로 표현한다.

미분

➢ 평균변화율

평균변화율은 함수 y = f(x)가 있을 때 $\dfrac{y\text{의 증가량}}{x\text{의 증가량}}$ $\dfrac{\Delta y}{\Delta x}$

예를 들어 f (x) = x^2인 경우, x가 1에서 5까지 증가했을 때 평균변화율은?

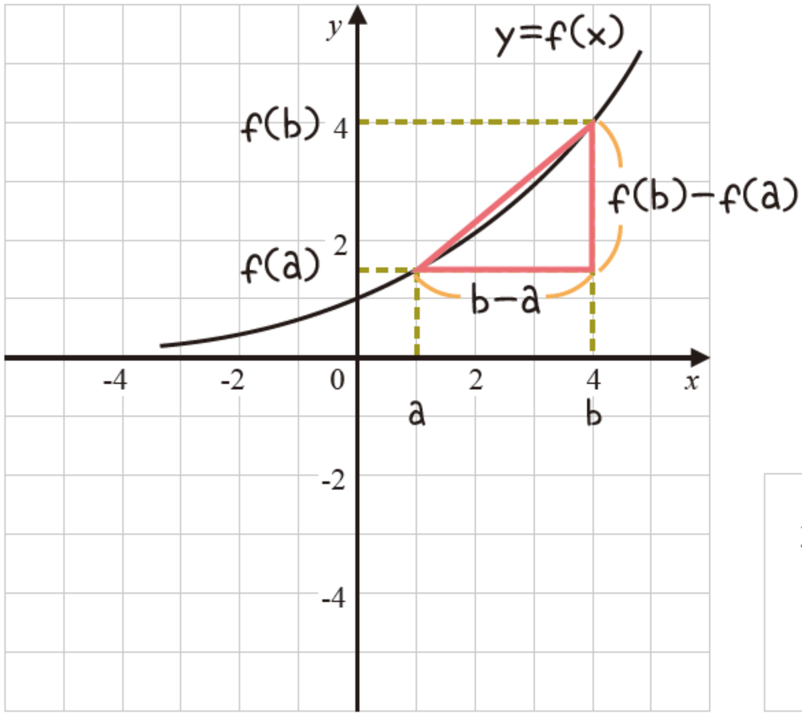
x의 증가량: Δx = 5 - 1 = 4

y의 증가량: Δy = f(5) - f(1)

$$\frac{\Delta y}{\Delta x} = \frac{f(5)-f(1)}{5-1} = \frac{25-1}{4} = 6$$

미분

➢ 평균변화율



x가 a에서 b로 변할 때 평균변화율

$$\frac{\Delta y}{\Delta x} = \frac{f(b) - f(a)}{b - a}$$

미분

➢ 평균변화율

또 다른 방식으로 평균변화율을 표현할 수도 있는데, Δx를 h로 치환하여 표현하는 것이다.

$$\frac{\Delta y}{\Delta x} = \frac{f(b)-f(a)}{b-a}$$ 에서

$$\frac{\Delta y}{\Delta x} = \frac{f(b)-f(a)}{b-a} = \frac{f(a+h)-f(a)}{h}$$

미분

➢ 평균변화율

연습 문제

함수 $f(x)=x^2$에 대해 x 값이 1에서 k까지 변할 때 평균변화율은 15입니다. 그렇다면 k 상수 값은 얼마일까요?

문제 풀이

$$\frac{\Delta y}{\Delta x}=\frac{f(k)-f(1)}{k-1}=\frac{k^2-1}{k-1}=\frac{(k+1)(k-1)}{k-1}=k+1=15$$ 이므로,

$k=14$입니다.

미분

평균변화율 코드 구현

```python
def average(a,b):
    m = max(a,b)
    n = min(a,b)
    x = Symbol('x')

    fx = 2 * x ** 2 + 4 * x + 7
    fb = fx.subs(x, m)
    fa = fx.subs(x, n)

    result = (fb - fa) / (m - n)
    return result

x1 = 0
x2 = 2

# 두점 (0,7) 와 (2,23) 의 평균 변화률은?
print(average(x1,x2))
```

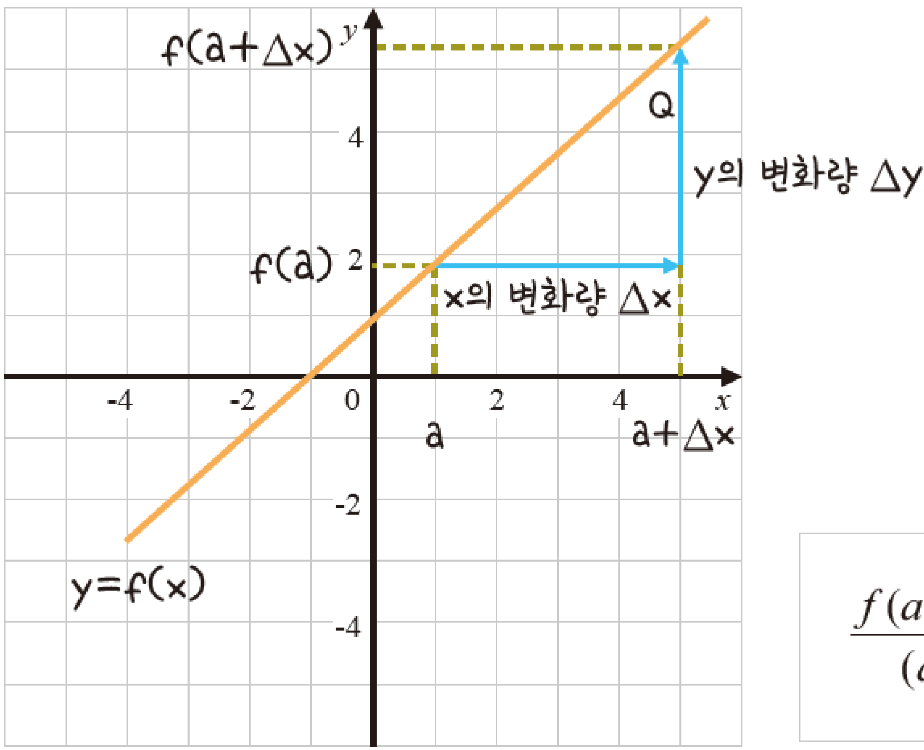
```
8
```

미분

➢ 미분계수(순간변화율)

- 미분계수는 x의 증가량이 0으로 가까이 갈 때 평균변화율

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$



$$\frac{f(a+\Delta x)-f(a)}{(a+\Delta x)-a}=\frac{f(a+\Delta x)-f(a)}{\Delta x}$$

미분

➢ **미분계수(순간변화율)**

- 이때 x의 변화량(Δx)을 0으로 가까이 보낸 것이 미분계수

$$f'(a)$$

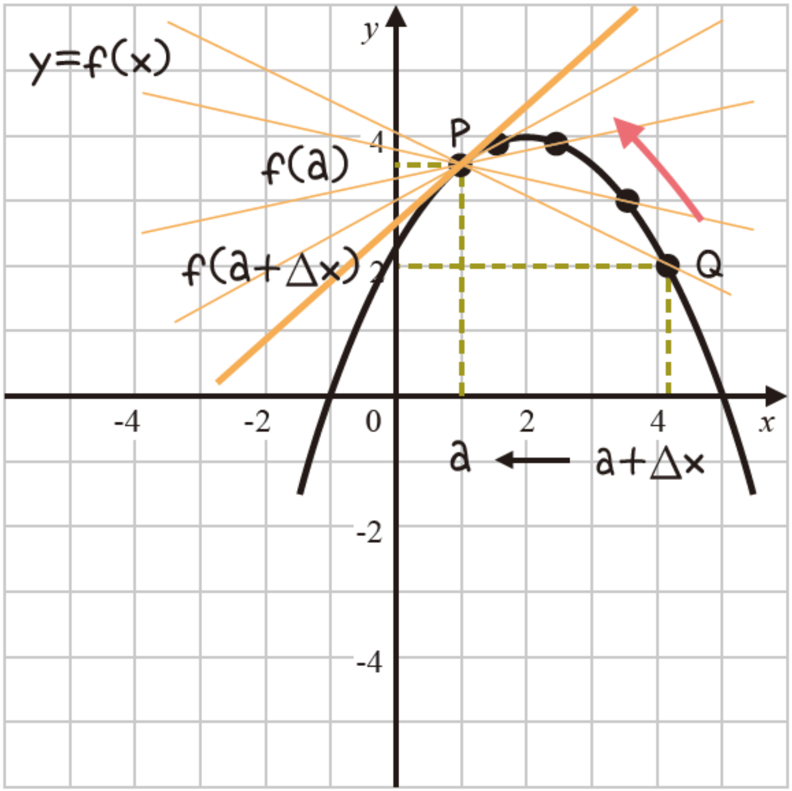
- 보통 Δx를 h로 표현할 때가 많으므로 f'(a)는 다음과 같이 나타낼 수 있음

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

$$= \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

미분

미분계수의 기하학적 의미



미분계수란 직선이 반시계 방향으로 돌아가면서 a + Δx가 a로 다가가면 직선 위의 점 P에서 기울기를 구하는 것과 같다.

함수 $y=f(x)$가 있을 때 $(a, f(a))$에서 접선의 기울기

미분

➢ 미분계수(순간변화율)

연습 문제

함수 $f(x)=x^2+3x$에 대해 x 값이 1에서 5로 변할 때 평균변화율과 $x=k$에서의 순간변화율이 같을 때 k 상수 값을 구하세요.

미분

➢ 미분계수(순간변화율)

또 $x = k$에서 순간변화율은 다음과 같습니다.

$$
\begin{aligned}
f'(a) &= \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} \\
&= \lim_{h\to 0} \frac{\{(k+h)^2+3(k+h)\}-(k^2+3k)}{h} \\
&= \lim_{h\to 0} \frac{k^2+2kh+h^2+3k+3h-k^2-3k}{h} \\
&= \lim_{h\to 0} \frac{2kh+h^2+3h}{h} \\
&= \lim_{h\to 0} \frac{h(2k+h+3)}{h} \\
&= 2k+h+3=9
\end{aligned}
$$

따라서 $h = 0$이고, $2k = 6$이므로 $k = 3$입니다.

미분

➢ 도함수

- 도함수란 함수 f(x)를 미분하여 얻은 함수 f'(x)임
- 예로 함수 f(x) = x^2 + x에서 f(1), f(2), f(3)을 구해보자

  - $$f'(1) = \lim_{h \to 0} \frac{f(1+h) - f(1)}{h}$$
  - $$f'(2) = \lim_{h \to 0} \frac{f(2+h) - f(2)}{h}$$
  - $$f'(3) = \lim_{h \to 0} \frac{f(3+h) - f(3)}{h}$$

미분

➢ 도함수

- 도함수를 이용하면 매번 계산의 번거로움을 줄이고자 미분계수를 함수처럼 생각해서 구할 수 있다.

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

다음은 도함수의 여러 다른 표기법

$$y', f'(x), \frac{dy}{dx}, \frac{d}{dx} f(x)$$

미분

표 8-1 도함수 기호의 차이

표현 방법	y'	$\frac{\triangle y}{\triangle x}$
읽는 방법	와이 프라임	디와이 디엑스
의미	y를 미분(y의 순간변화율)	y를 x에 대해 미분(x에 대한 y의 순간변화율)
수학자	뉴턴	라이프니츠
연구	한 물체의 운동 방향 연구	특정한 두 변수의 관계 연구

미분

➢ 도함수

연습 문제

함수 $f(x) = 2x^2 - 1$의 도함수를 구하세요. 또 이 도함수를 이용하여 $f(x)$에서 $x = 6$에서의 미분계수를 구하세요.

미분

파이썬 코드로 구현

```python
from sympy import Derivative, Symbol, diff
```

$$2 \times x^2 + 4 \times x + 7$$

```python
x = Symbol('x')
fx = 2 * x ** 2 + 4 * x + 7

fprime = Derivative(fx, x).doit()
n = fprime.subs({x: 3})

print(fx)
print(fprime)
print("fx에서 x = 3 에서의 순간변화율(미분계수는) ", n , "입니다")
```

미분

파이썬 코드로 구현

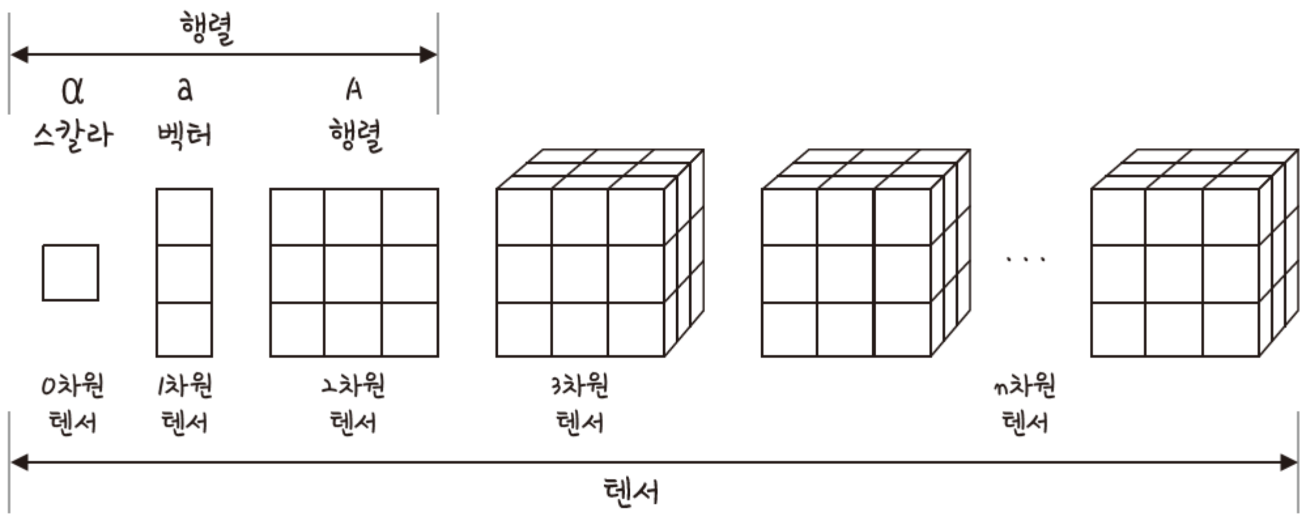
```python
from sympy import Derivative, Symbol, diff
```

```python
x = Symbol('x')
fx = 2 * x ** 2 + 4 * x + 7

fx = 2*x**2 + 4*x + 7

fprime2 = diff(fx, x)

n = fprime2.subs({x: 3})

print("fx에서 x = 3 에서의 순간변화율(미분계수는) ", n , "입니다")
```

선형대수학

➢ 선형대수(Linear algebra)란?

- 선형대수(linear algebra)는 데이터 분석에 필요한 각종 계산을 돕는 학문
- 선형대수에서 다루는 데이터는 개수나 형태에 따라 크게 스칼라(scalar), 벡터(vector), 행렬(matrix), 텐서(tensor) 유형으로 나눌 수 있다.
- 스칼라 : 숫자 하나로 구성된 데이터
- 벡터 : 여러 숫자로 구성된 데이터 레코드(data record)
- 행렬 : 벡터, 즉 데이터 레코드가 여럿인 데이터 집합
- 텐서 : 크기가 같은 행렬이 여러 개 있는 형태



선형대수학

➢ 선형대수에서 스칼라란?

- 스칼라는 숫자 하나만으로 된 데이터
- 예를 들어 IT 직군에 있는 한 사람의 평균 급여를 측정하면 숫자가 하나만 나올 것이 데이터가 스칼라
- 스칼라는 보통 x처럼 알파벳 소문자로 표시하며, 실수(real number)인 숫자 중 하나이므로 실수 집합 R의 원소라는 의미에서 다음과 같이 표기한다.

$$x \in R$$

선형대수학

➢ 선형대수에서 벡터란

- 벡터는 다음과 같이 세로줄(열(column))은 하나고, 가로줄(행(row))은 여러 개인 형태로 위에서 아래로 써서 표기해야 한다.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

- 벡터 하나를 이루는 데이터 개수가n 개이면 이 벡터를 n-차원 벡터(n-dimensional vector)라고 하며, 다음과 같이 표기함

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x \in R^n$$

선형대수학

➢ 선형대수에서 벡터란

- 실제로 앞서 예로 든 '이직 유무'를 알아내고자 '업무 만족도', '전년도 평가', '평균 급여', '연봉'이라는 네 항목을 측정했으므로 4차원 벡터라고 하며, 다음과 같이

$$x \in R^4$$

업무 만족도	전년도 평가	프로젝트 수	평균 급여	업무 시간	이직 유무	직군	연봉
0.48	0.43	3	96	3	0	support	low
0.5	0.58	4	97	3	1	sales	low
0.79	0.61	5	98	4	1	marketing	medium
0.34	0.67	5	99	2	1	IT	low
0.45	0.79	5	112	6	0	accounting	high
0.28	0.89	4	113	6	0	management	low
0.67	0.36	4	102	4	1	sales	medium
0.78	0.44	3	103	4	0	IT	medium
0.47	0.57	3	109	4	1	support	low
0.27	0.48	3	105	6	1	IT	medium

선형대수학

➢ 선형대수에서 벡터란

- 데이터 레코드 하나를 단독으로 벡터로 나타낼 때는 열 하나로 표시한다.
- 인공지능에서는 행을 특성 행렬(feature matrix)이라고도 한다.
- 행렬의 크기를 수식으로 표시할 때는 다음과 같이 '행의 크기×열의 크기'로 나타낸다.
- 예를 들어 행이 여섯 개고 열이 네 개이므로 다음과 같이 표현함

$$x \in R^{6\times4}$$

선형대수학

➢ 선형대수에서 텐서란?

텐서: 3차원 이상의 배열

- 텐서는 크기가 같은 행렬이 여러 개 같이 묶여 있는 것을 의미한다.
- 데이터 과학(data science) 분야에서는 흔히 다차원 배열을 텐서라고 한다.

선형대수학

➢ 인공지능에서는 왜 벡터를 사용하는가?

표 10-4 HR 테이블

이직 유무	평균 급여	업무 시간	직군	연봉
0	96	3	support	low
1	97	3	sales	low
1	98	4	marketing	medium
1	99	2	IT	low
0	112	6	accounting	high

선형대수학

➢ 인공지능에서는 왜 벡터를 사용하는가?

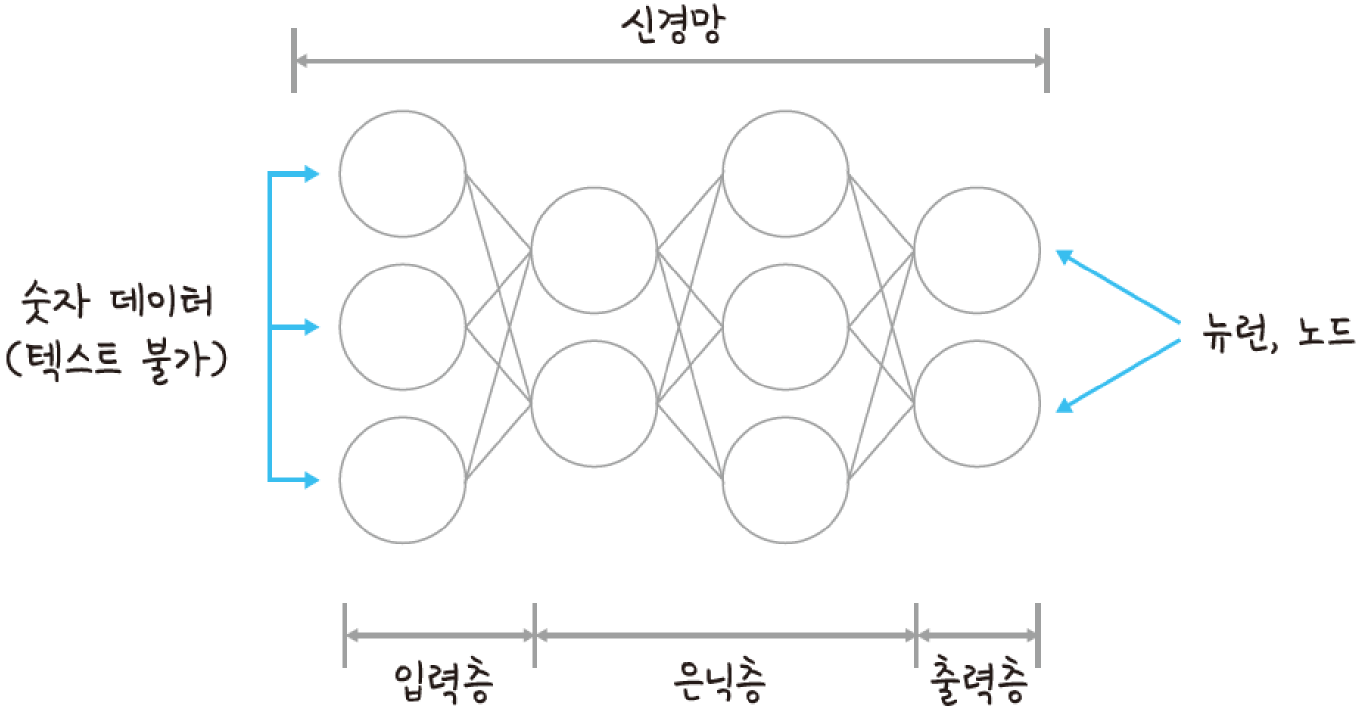
표 10-5 데이터의 숫자 표현 결과

이직 유무	평균 급여	업무 시간	직군	연봉
0	96	3	1	1
1	97	3	2	1
1	98	4	3	2
1	99	2	4	1
0	112	6	5	3

선형대수학

➢ 인공지능에서는 왜 벡터를 사용하는가?

그림 10-17 인공신경망



선형대수학

표 10-1 HR 테이블

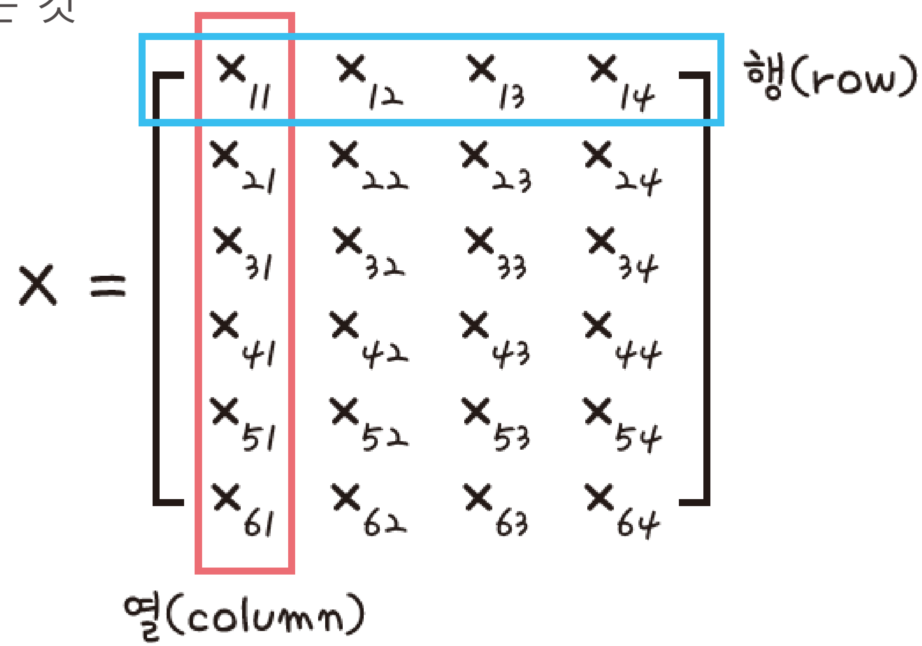
업무 만족도	전년도 평가	프로젝트 수	평균 급여	업무 시간	이직 유무	직군	연봉
0.48	0.43	3	96	3	0	support	low
0.5	0.58	4	97	3	1	sales	low
0.79	0.61	5	98	4	1	marketing	medium
0.34	0.67	5	99	2	1	IT	low
0.45	0.79	5	112	6	0	accounting	high
0.28	0.89	4	113	6	0	management	low
0.67	0.36	4	102	4	1	sales	medium
0.78	0.44	3	103	4	0	IT	medium
0.47	0.57	3	109	4	1	support	low
0.27	0.48	3	105	6	1	IT	medium

선형대수학

➢ 선형대수에서 행렬이란?

행렬: 2차원 배열

- 행렬은 복수 차원을 가지는 데이터 레코드가 여러 개 있을 때, 이 데이터를 합쳐 표기한 것이다.
- 예를 들어 앞의 **HR** 테이블에서 이직에 성공한 사람들(‘이직 유무’가 ‘1’인 사람들이 총 6명)의 ‘업무 만족도’, ‘전년도 평가’, ‘평균 급여’, ‘연봉’ 데이터를 추출했기 때문에 이렇게 4차원 데이터가 여섯 개 있는 것이다.
- 즉, 4 × 6 = 24개의 실수 숫자가 있는 것
- 행렬은 다음과 같이 보통 X 등 알파벳 대문자로 표시한다.

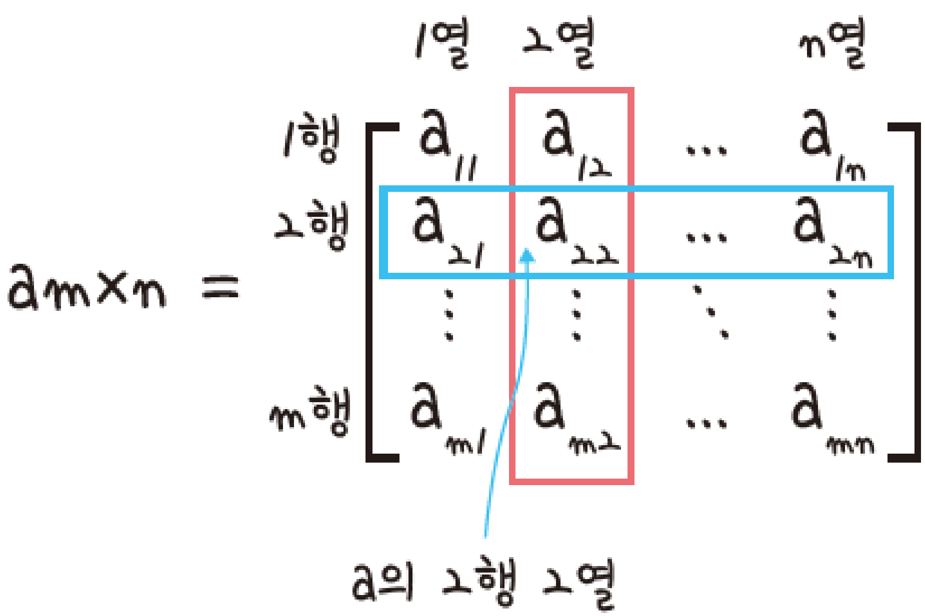


선형대수학

➢ 행렬 더 알아보기

- 행렬의 표현은 상수나 변수를 직사각형 꼴로 배열한 것으로, 엑셀의 행과 열을 생각하면 쉬움

그림 10-4 행렬의 표현



행렬의 구성 요소

- 성분: 행렬을 이루는 각각의 상수나 변수
- 행: 성분의 가로 배열
- 열: 성분의 세로 배열
- m×n 행렬: 행 m개와 열 n개로 된 행렬

선형대수학

➢ 행렬로 표현하기

- 예를 들어 학생 A, B가 가진 과목별 참고서 수량이 표 10-3과 같다고 하자

표 10-3 과목별 참고서 수량

학생 \ 참고서	수학	영어	국어
A	5	2	2
B	1	7	3

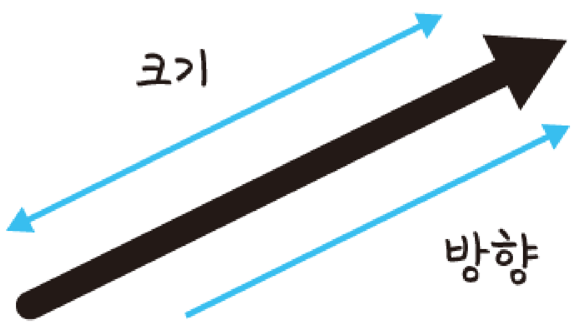
$$\begin{pmatrix} 5 & 2 & 2 \\ 1 & 7 & 3 \end{pmatrix}$$

행렬 안에 표현된 각각의 성분 5, 2, 2, 1, 7, 3을 **행렬의 성분**이라고 함

➢ 벡터의 기하학적 정의

- 벡터는 다음과 같이 '크기(magnitude)와 방향(direction)을 가진 물리량'임
- 수치와 연관되지 않고 공간상 '임의의 점(시작점)에서 임의의 점(끝점)을 연결한 화살표'이기 때문에 벡터는 수치적인 개념보다 기하학적 혹은 시각적으로 이해해야 한다.

그림 10-7 벡터의 개념



예시

비행체가 $3km$ 속도로 동북쪽 방향으로 운행 중입니다.

크기: $3km$ 속도

방향: 동북쪽

선형대수학

대수적 표기

- 벡터는 다음과 같이 소문자 볼드체로 표시함

$$\mathbf{x} = (x_1, x_2, \cdots, x_n) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T \quad \text{수식 10.1}$$

(T는 Transpose의 머리글자로 전치행렬을 의미합니다.

전치행렬은 열과 행을 바꾸어서 표현합니다.)

(1) $\mathbf{x} = [3 \;\; 2]$

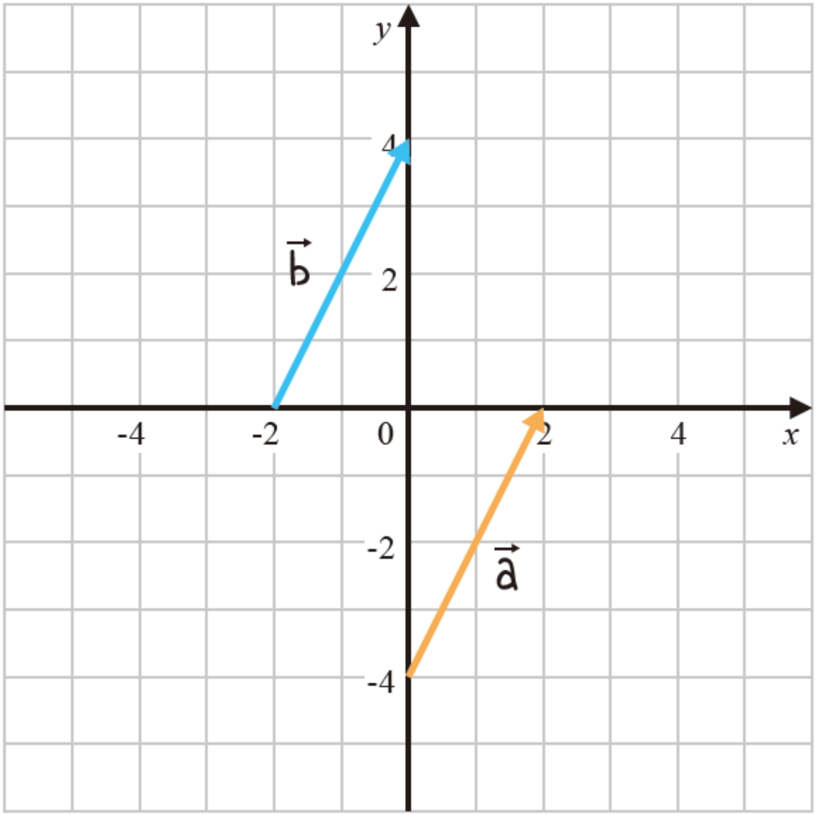
(2) $\mathbf{x}^T = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

선형대수학

➢ 특수한 벡터

단위 벡터

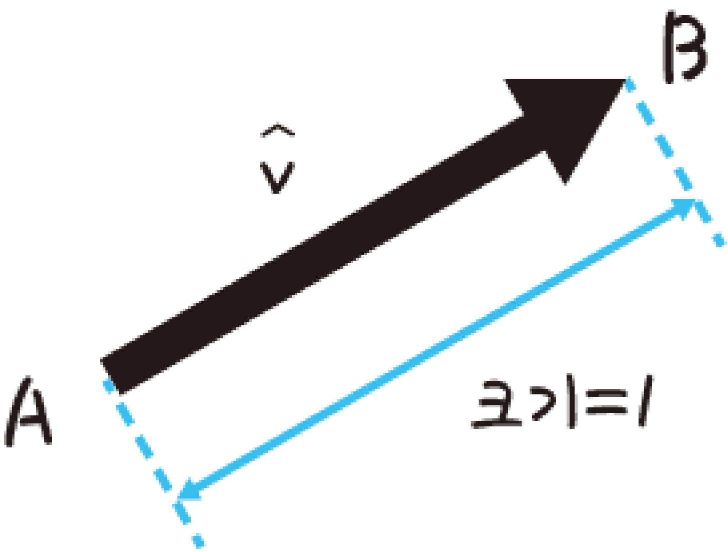
- 단위 벡터는 **길이가 1인 벡터를 의미**하며, 길이가 1이기 때문에 방향 성분만 나타낸다. 단위 벡터는 크기가 1이 되도록 늘리거나 줄인 벡터를 의미한다.
- 즉, 기하학적으로는 같은 벡터이지만 대수학적으로 벡터 $\vec{a}$가 벡터 $\vec{b}$는 의미가 다름



선형대수학

➢ 특수한 벡터

단위 벡터 표기법



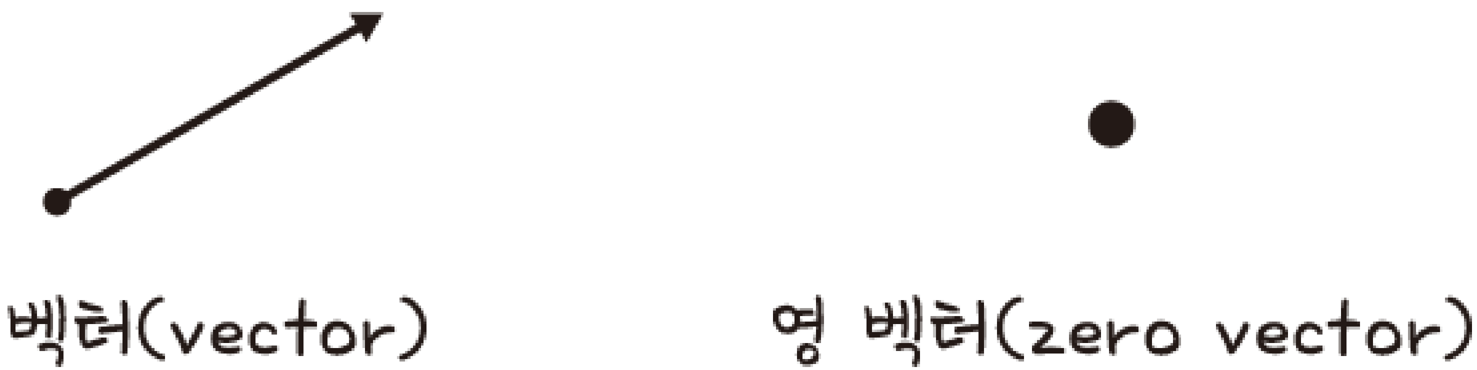
$$\frac{\text{벡터 }\vec{v}}{\text{벡터 }\vec{v}\text{의 크기}} = \frac{\vec{v}}{|\vec{v}|} = \frac{\vec{v}}{||\vec{v}||}$$

선형대수학

➢ 특수한 벡터

영 벡터

- 영 벡터는 모든 성분이 0으로 구성된 벡터
- 크기(길이)가 0이면서 시작점과 끝점이 같은 벡터를 영 벡터라고 한다.



$$\vec{0} \text{ 혹은 } \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \end{bmatrix}$$

선형대수학

➢ 특수한 벡터

영 벡터의 특징

- 크기는 0이지만 방향은 모든 방향임
- 방향이 모든 방향이기 때문에 다음 특징이 있음

  - 어떤 벡터와도 평행임
  - 어떤 벡터와도 같은 직선 위에 있음
  - 어떤 벡터와도 수직임
  - 덧셈과 뺄셈의 항등원임

선형대수학

➢ 벡터 공간과 부분 공간의 기저

항등행렬

- 항등행렬은 중심 대각선을 제외하고 모든 항목이 0인 행렬로, 중심 대각선의 항목은 모두 1임

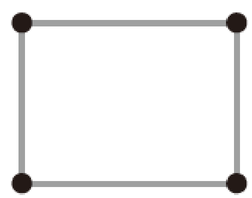
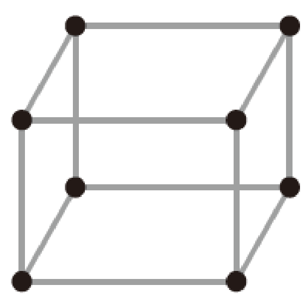
$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$MI = IM = M$$

선형대수학

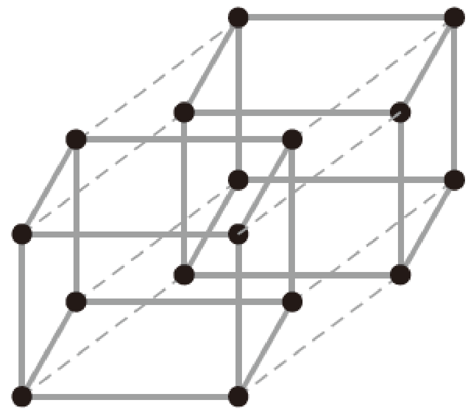
차원

- 선형대수학적 차원에서 차원은 기저 벡터의 개수를 의미

차원	표현		설명
0차원	점		점으로 표현합니다.
1차원	선		선으로 표현하며, 길이를 측정하는 데 사용합니다.
2차원	면		면으로 표현하며, 넓이를 측정하는 데 사용합니다.
3차원	입체		입체로 표현하며, 부피를 측정하는 데 사용합니다.

선형대수학

차원

차원	표현		설명
4차원	하이퍼큐브 (hypercube)		초입체로 표현하며, 초부피를 측정하는 데 사용합니다.

선형대수학

랭크

- 랭크(rank)는 행렬 A에서 선형 독립인 행 혹은 열의 개수를 의미
- 예를 들어 행렬 A를 1행과 2행의 선형 조합으로 3행을 만든다

$$A = \begin{pmatrix} 3 & 1 & 3 \\ 4 & 0 & 3 \\ -1 & 1 & 0 \end{pmatrix}$$

선형 독립인 행 개수는 2, 행렬 A의 랭크는 2

R^m의 벡터 n개 $\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix}, \cdots, \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix}$ 중 선형 독립인 벡터 개수

$rk(A)$ 혹은 $rank(A)$

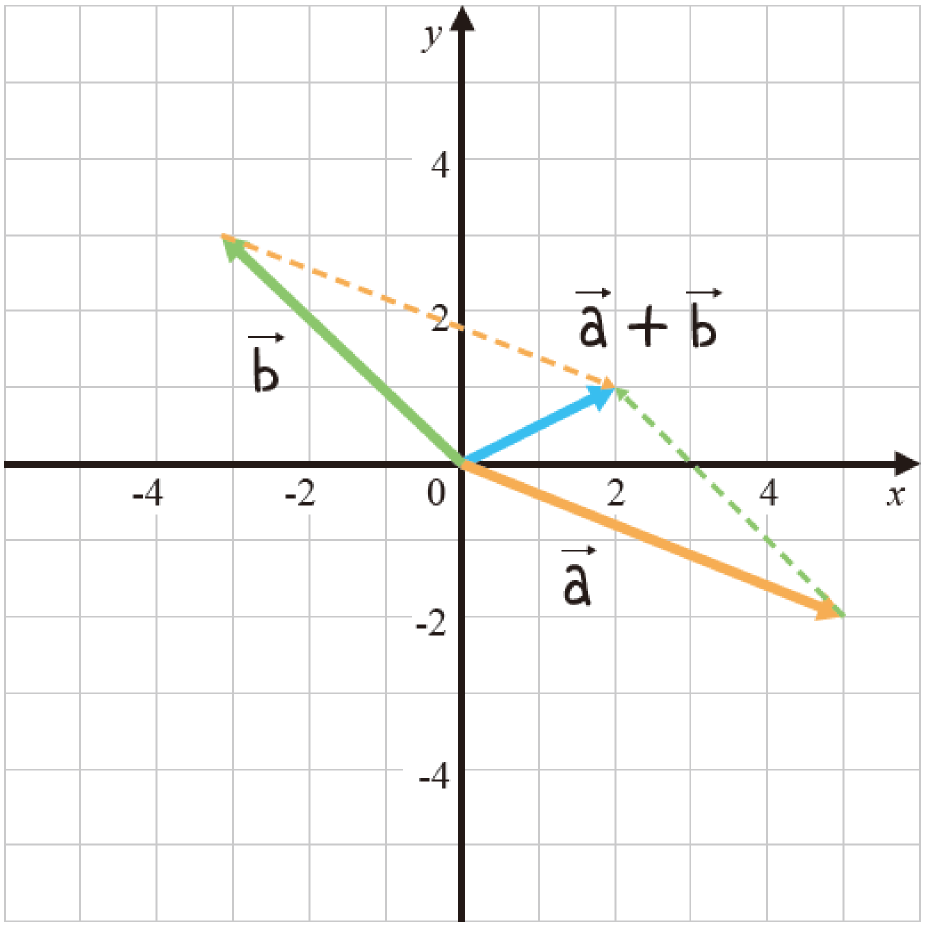
선형대수학

➢ 벡터의 덧셈

$$\vec{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \ \vec{w} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{일 때, } \vec{v} + \vec{w} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \qquad \text{수식 10.9}$$

평행사변형 법칙	삼각형 법칙
$\vec{a}$와 $\vec{b}$의 시작점이 일치할 때 평행사변형의 대각선은 두 벡터의 합을 의미합니다. 이때 $\lvert\vec{a}+\vec{b}\rvert$는 두 벡터 합의 크기가 됩니다.	$\vec{a}$의 끝점과 $\vec{b}$의 시작점을 연결할 때 시작점과 끝점을 연결한 벡터는 두 벡터의 합을 의미합니다. 이때 $\lvert\vec{a}+\vec{b}\rvert$는 두 벡터 합의 크기가 됩니다.

선형대수학



선형대수학

벡터 덧셈의 성질

(1) 교환 법칙: A + B = B + A

예시 $\vec{v}=(12,\ 17),\ \vec{w}=(2,\ -5)$일 때,

$$\vec{v}+\vec{w}=(12,\ 17)+(2,\ -5)=(14,\ 12)$$

$$\vec{w}+\vec{v}=(2,\ -5)+(12,\ 17)=(14,\ 12)$$

(2) 결합 법칙: (A + B) + C = A + (B + C)

예시 $\vec{v}=(12,\ 17),\ \vec{w}=(2,\ -5),\ \vec{u}=(4,\ -3)$일 때,

$$(\vec{v}+\vec{w})+\vec{u}=((12,\ 17)+(2,\ -5))+(4,\ -3)=(18,\ 9)$$

$$\vec{v}+(\vec{w}+\vec{u})=(12,\ 17)+((2,\ -5)+(4,\ -3))=(18,\ 9)$$

(3) 벡터 덧셈의 항등원이 존재: A + 0 = A
(4) 벡터 덧셈의 역원이 존재: A + (-A) = 0

선형대수학

➢ 벡터의 곱셈

연습 문제

다음 문제에서 벡터의 내적을 계산하세요.

(1) $\begin{bmatrix} 3 \\ -6 \end{bmatrix}\begin{bmatrix} -7 \\ 9 \end{bmatrix}$

(2) $\begin{bmatrix} -3 & 4 & 7 \end{bmatrix}\begin{bmatrix} -4 \\ -9 \\ 5 \end{bmatrix}$

(3) $\begin{bmatrix} 3 \\ -11 \\ 7 \end{bmatrix}\begin{bmatrix} -4 \\ -13 \\ 9 \end{bmatrix}$

선형대수학

➢ 벡터의 곱셈

문제 풀이

(1) $\begin{bmatrix} 3 & -6 \end{bmatrix}^T \begin{bmatrix} -7 \\ 9 \end{bmatrix} = (3 \cdot (-7)) + ((-6) \cdot 9)$

$= ((-21) + (-54)) = -75$

선형대수학

➢ 벡터의 곱셈

(2)

$$\begin{bmatrix} -3 & 4 & 7 \end{bmatrix} \begin{bmatrix} -4 \\ -9 \\ 5 \end{bmatrix} = ((-3)\cdot(-4)) + (4\cdot(-9)) + (7\cdot 5)$$

$$= (12 + (-36) + 35) = 11$$

선형대수학

➢ **벡터의 곱셈**

(3)

$$\begin{bmatrix} 3 & -11 & 7 \end{bmatrix}^T \begin{bmatrix} -4 \\ -13 \\ 9 \end{bmatrix} = (3 \cdot (-4)) + ((-11) \cdot (-13)) + (7 \cdot 9)$$

$$= ((-12) + 143 + 63) = 194$$

선형대수학

➢ 선형대수학 파이썬 구현

Numpy란? Numeric + Python = Numpy(넘파이, 늄파이)

수학 및 과학 연산을 위한 파이썬 패키지

관련 사이트 : http://www.numpy.org

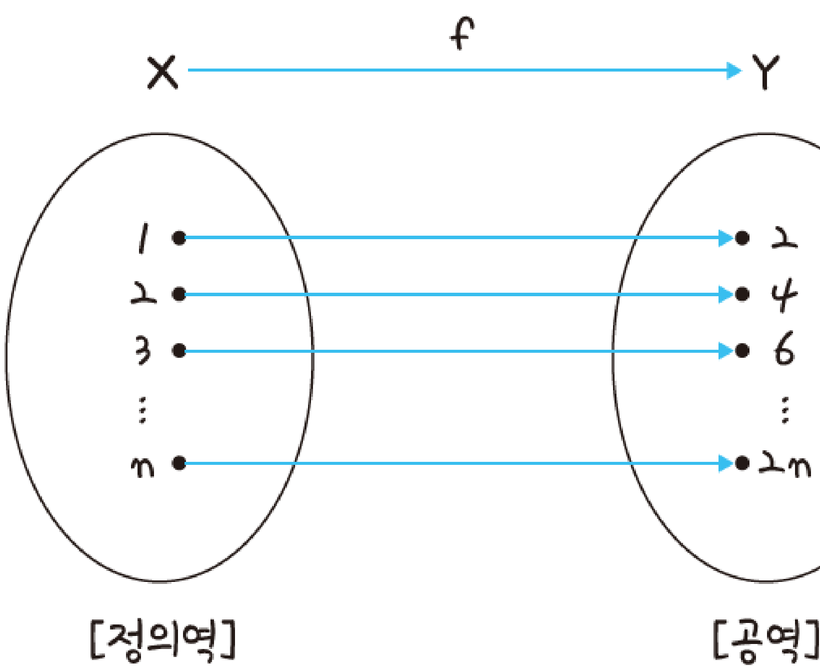
```
arr = np.arange(1,21).reshape(4,5)
arr
```

```
array([[ 1,  2,  3,  4,  5],
       [ 6,  7,  8,  9, 10],
       [11, 12, 13, 14, 15],
       [16, 17, 18, 19, 20]])
```

➢ 수열의 개념 및 종류

수열(Sequence) 란?

- 수열(sequence)은 규칙성이 있는 숫자의 나열
- 수열을 수학적으로 정의하면, 자연수를 정의역으로 갖는 함수나 그 함수의 결과로 얻는 원소들을 나열하는 것이다.



함수 f: X → Y에서 함수 값 f(1), f(2), f(3), f(4)를 수열의 항이라고 한다.

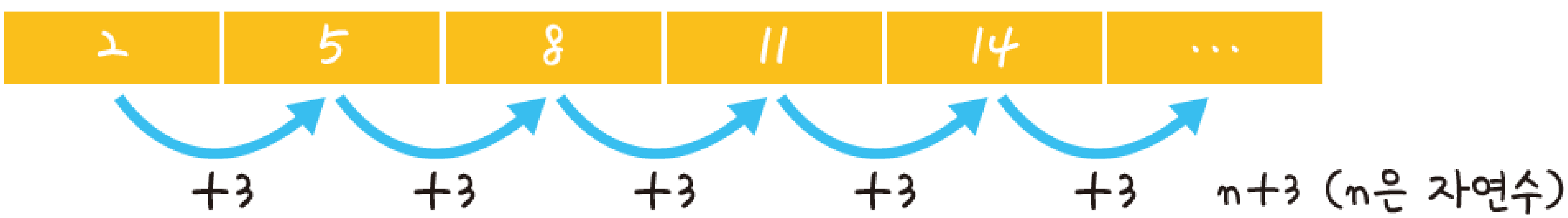
확률과 통계

➢ 수열의 개념 및 종류

- 수열은 2, 4, 6, ..., n이 되고 집합 기호를 사용하여 $\{a_n\}$으로 표현한다.

$$\{2,\ 4,\ 6,\ \cdots,\ n\} \Leftrightarrow a_n = 2n$$

- 등차수열(arithmetic sequence)은 연속하는 두 항의 차이가 모두 일정한 수열임
두 항의 차이는 이 수열의 모든 항에 대해 공통적으로 나타나는 차이이므로
공차(common difference)라고 한다.

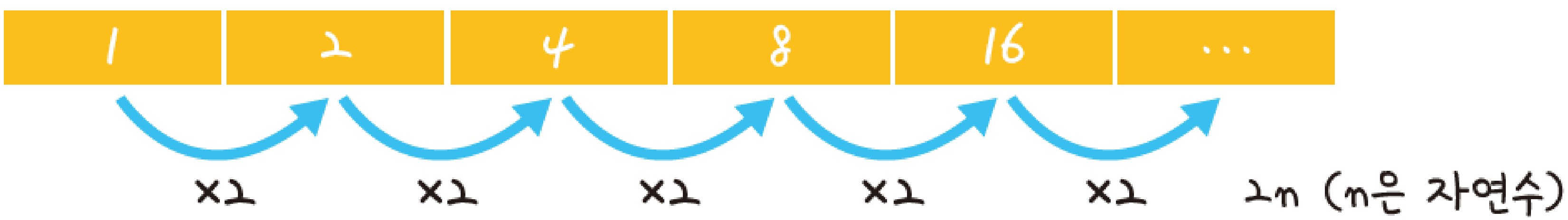


확률과 통계

➢ 수열의 개념 및 종류

- 등비수열(geometric sequence)은 수열의 각 항이 그 앞의 항에 일정한 수를 곱한 것으로 이루어진 수열임
- 첫 항은 0이 되어서는 안 되며, 곱하는 일정한 수를 공비(common ratio)라고 한다.

그림 13-3 등비수열



구분	등차수열	등비수열
정의	어떤 수에 차례로 일정한 수를 더해서 얻는 수열	어떤 수에 일정한 수를 곱해서 얻는 수열
일반항	$a_n = a + (n-1)d$ (a: 수열, d: 공차)	$a_n = ar^{n-1}$ (a: 수열, r: 공비)

확률과 통계

➢ 순열(permutation)

순열이란

- 서로 다른 n개에서 서로 다른 r개를 선택하여 일렬로 나열하는 것을 n개에서 r개를 택한 순열이라고 하며, 이 순열의 수를 다음과 같이 표기함

$$_{n}P_{r}$$

- 서로 다른 n개에서 서로 다른 r개를 택하는 순열의 수는 다음과 같이 표현함

$$_{n}P_{r} = \underbrace{n(n-1)(n-2)\cdots(n-r+1)}_{r\text{개}} \ (\text{단 } 0 < r \le n)$$

확률과 통계

➢ 순열

- n 팩토리얼(factorial) :
  1부터 n까지 자연수를 차례로 곱한 것을 n의 계승(혹은 팩토리얼)이라고 한다

$$n! = n(n-1)(n-2) \quad \cdots \quad \cdot 3 \cdot 2 \cdot 1$$

$$_{n}p_{r} = \frac{n!}{(n-r)!}$$

- 예를 들어 1, 2, 3, 4, 5처럼 숫자 다섯 개에서 세 개를 선택하여 세 자리 자연수를 만드는 방법의 수를 구해 보자

(1) 일반 수식을 이용한 경우: $_{5}P_{3} = 5 \times 4 \times 3 = 60$

(2) 팩토리얼을 이용한 경우: $_{5}P_{3} = \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = 60$

확률과 통계

➢ 순열

연습 문제

트럼프 카드가 열 장 있습니다. 열 장 중 임의로 세 장을 뽑을 수 있는 방법의 수를 구하세요.

확률과 통계

➢ 순열

중복순열 (permutation with repetition)

- 순열이 서로 다른 n개에서 중복을 허락하지 않고 r개를 일렬로 나열하는 수였다면 중복순열은 중복을 허락하고 r개를 일렬로 나열하는 수를 의미

$$ {}_{n}\Pi_{r} = n^{r} $$

확률과 통계

➢ 순열

연습 문제

(1) 1, 2, 3, 4, 5, 6을 사용하여 중복을 허락하지 않고 세 자리 자연수를 만드는 방법의 수를 구하세요.

(2) 1, 2, 3, 4, 5, 6을 사용하여 중복을 허락하고 세 자리 자연수를 만드는 방법의 수를 구하세요.

확률과 통계

➢ 조합

조합이란 : 서로 다른 n개에서 순서를 생각하지 않고 r개를 선택하는 것을 n개에서 r개를 택한 조합이라고 하며, 이 조합의 수를 기호로 다음과 같이 표현함

$$ {}_{n}C_{r} $$

$$ {}_{n}C_{r} = \frac{{}_{n}P_{r}}{r!} = \frac{\dfrac{n!}{(n-r)!}}{\dfrac{r!}{1}} $$

곱한 후 분모

곱한 후 분자

$$ = \frac{n!}{r!(n-r)!} \quad (\text{단 } 0 < r \le n) $$

확률과 통계

➢ 조합

- 학생 100명 중에서 98명을 선택하는 경우의 수를 생각해 보자

$$
{}_{100}C_{98} = {}_{100}C_{2} = \frac{{}_{n}P_{r}}{r!} = \frac{\frac{100!}{(100-2)!}}{2!}
$$

$$
= \frac{\frac{100!}{98!}}{2!} = \frac{100 \times 99}{2 \times 1} = 4950
$$

확률과 통계

➢ 조합

연습 문제

한 반의 학생 수가 40명입니다.

(1) 반장 1명, 부반장 1명을 뽑는 경우의 수를 구하세요.

(2) 주번 3명을 뽑는 경우의 수를 구하세요.

확률과 통계

➢ 순열과 조합의 비교

구분	순열	조합
순서와 위치	순서와 위치가 중요합니다.	순서와 위치가 중요하지 않습니다.
표현	'배열하다'로 표현($_nP_r$)	'뽑는다'로 표현($_nC_r$)
계산 방법	$_nP_r = \dfrac{n!}{(n-r)!}$	$_nC_r = \dfrac{n!}{r!(n-r)!}$
배열 방법	배열 방법을 정하지 않았습니다.	배열 방법을 한 가지로 정했습니다.

구분	순열	중복순열	조합	중복조합
순서	있음	있음	없음	없음
중복	불가능	가능	불가능	가능
표현	$_nP_r$	$_n\Pi_r = n^r$	$_nC_r$	$_nH_r$

확률과 통계

➢ 파이썬 코드로 구현

```
import numpy as np
```

```
# 등차 수열
n = 3
list(range(1,51,n))
```

```
[1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, 34, 37, 40, 43, 46, 49]
```

```
n = 3
np.arange(1,51,n)
```

```
array([ 1,  4,  7, 10, 13, 16, 19, 22, 25, 28, 31, 34, 37, 40, 43, 46, 49])
```

확률과 통계

➢ 파이썬 코드로 구현

```
# 등비 수열
n = 2 # 등비
s = 1  # 첫번째 항
result = [s]
for i in range(1,11):
    result.append(result[i-1]*n)
result
```

```
[1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024]
```

확률과 통계

➢ 파이썬 코드로 구현

```
# 등비 수열
n = 2 # 등비
s = 1  # 첫번째 항
result = [s]
for i in range(1,11):
    result.append(result[i-1]*n)
result
```

```
[1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024]
```

확률과 통계

➢ **파이썬 코드로 구현**

```
from itertools import permutations
```

```
# 카드 1, 2, 3 으로 구성할 수 있는 경우의 수
lists = [1, 2, 3]
print(list(permutations(lists, 2)))
```

```
[(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)]
```

```
from itertools import combinations
```

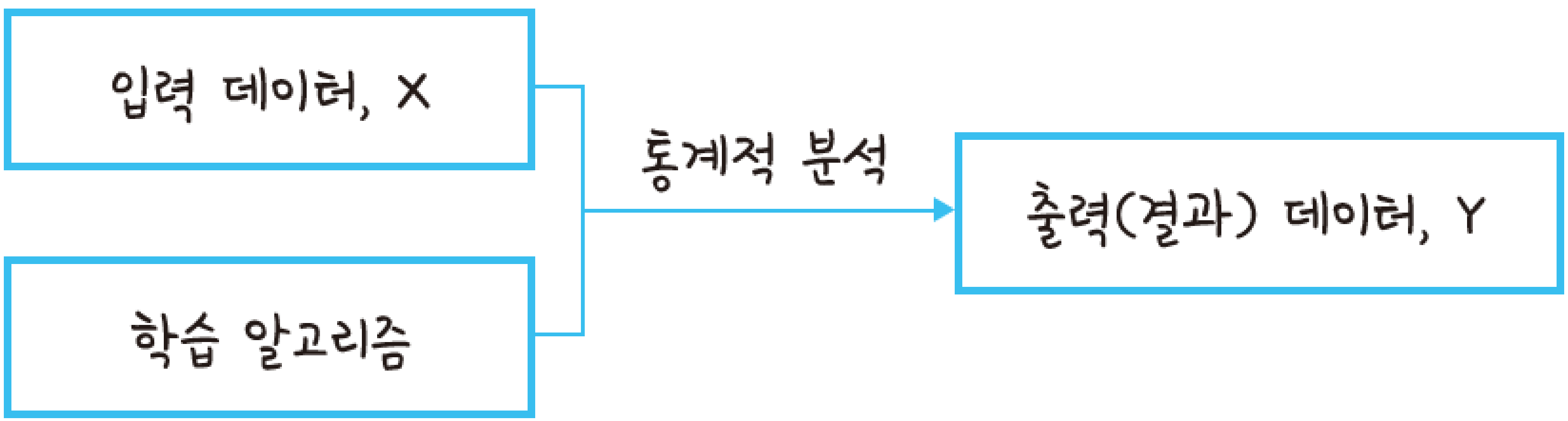
```
# 카드 3장에서 2장을 선택한다
cards = [1, 2, 3]
list(combinations(cards, 2))
```

```
[(1, 2), (1, 3), (2, 3)]
```

확률과 통계

➢ 확률과 인공지능의 관계

통계적 분석은 **독립변수(X)**와 수학적 모델을 입력으로 알려 주면 **종속변수(Y)**를 출력

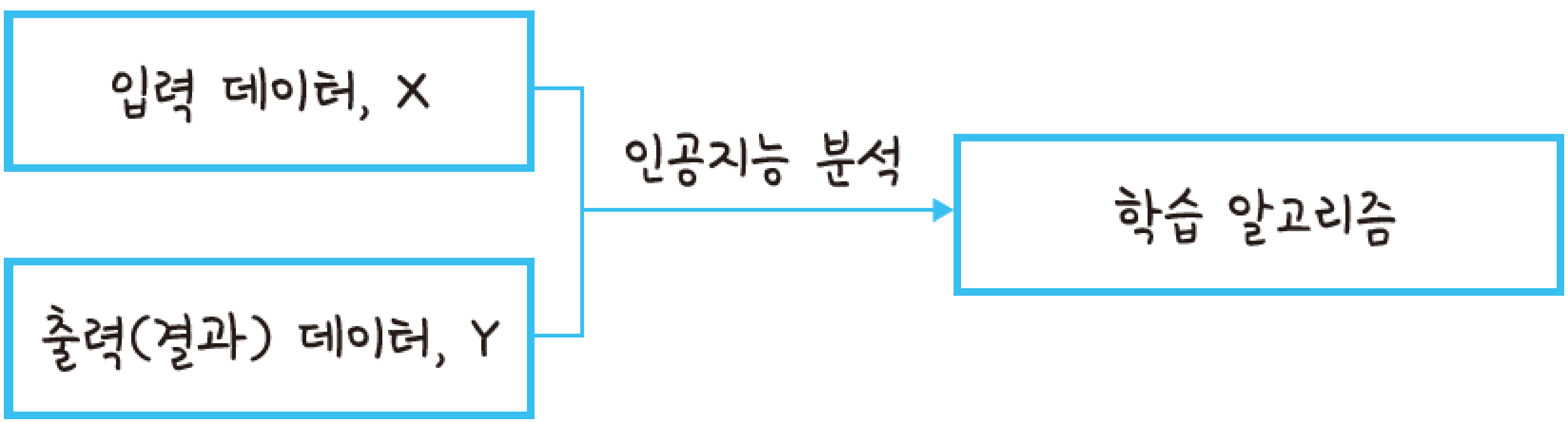


확률과 통계

➢ 확률과 인공지능의 관계

- 인공지능을 이용한 분석은 독립변수(X)와 종속변수(Y, label)를 알려 주면 컴퓨터 스스로 학습 모델을 만든다

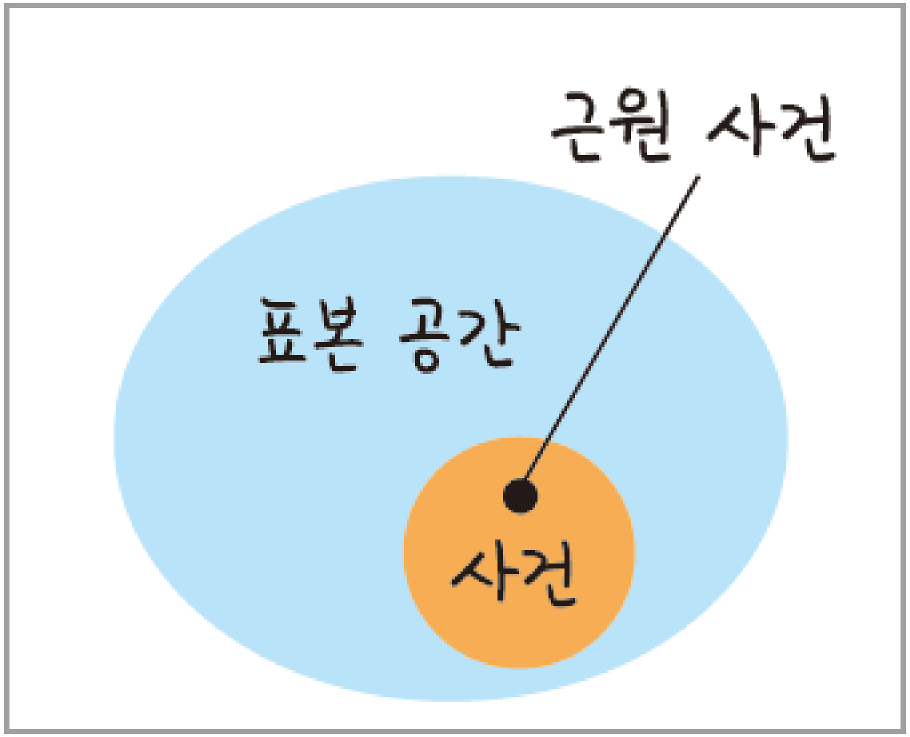
▼ 그림 14-2 인공지능 분석



확률과 통계

➢ 확률 기본 용어

▼ 그림 14-3 사건 및 표본 공간



확률과 통계

표 14-1 확률 기본 용어

용어	설명	표현
실험(trial)	동일한 조건에서 여러 번 반복할 수 있고 그 결과가 우연으로 결정되는 관찰이나 실험	
표본 공간 (sample space)	한 실험에서 나올 수 있는 모든 가능한 결과의 집합	Ω
근원사건 (elementary outcome)	표본 공간을 이루는 개개의 결과	$\omega_1, \omega_2, \cdots$
사건(event)	근원사건의 집합, 표본 공간의 부분 집합	
합사건	두 사건 A와 B의 합집합으로 표현할 수 있는 사건	$A \cup B$
곱사건	두 사건 A와 B의 교집합으로 표현할 수 있는 사건	$A \cap B$
여사건	사건 A가 일어나지 않는 사건	A^C
배반사건	사건 A와 B가 동시에 일어나지 않는 사건	$A \cap B = 0$

확률과 통계

➢ 확률 기본 용어

- 예를 들어 주사위를 1회 던졌을 때, 표본 공간과 근원사건을 알아보자

  - 표본 공간(Ω): Ω = {1, 2, 3, 4, 5, 6}
  - 근원사건(ω_1, ω_2, ...): 1, 2, 3, 4, 5, 6
  - 짝수가 나올 사건(A, B, ...): A = {2, 4, 6}

확률과 통계

➢ 확률 기본 용어

사건의 확률과 확률의 성질

- 어떤 실험에서 사건 A가 일어날 가능성을 수로 나타낸 것을 사건 A가 일어날 **확률**이라고 하며, 다음과 같이 표현한다.

$$P(A)$$

$$P(A) = \frac{A\text{가 일어나리라 예상되는 횟수}}{\text{전체 실험의 횟수}}$$

확률과 통계

➢ 확률 기본 용어

- 이것을 다음과 같이 좀 더 수학적으로 표현할 수 있는데, 이를 수학적 확률이라고 함

$$P(A) = \frac{n(A)}{n(S)}$$

- 확률은 다음 세 가지 성질을 갖음

(1) 임의의 사건 A에 대해 확률 P(A)는 $0 \le P(A) \le 1$임

(2) 반드시 일어나는 사건 S에 대해 확률 $P(S) = 1$임

(3) 절대로 일어나지 않는 사건 ø에 대해 $P(ø) = 0$임

- 즉, 확률의 최소는 0이고, 최대는 1이라고 정리할 수 있다

확률과 통계

➢ 독립사건과 종속사건

두 사건 A와 B에서 한 사건의 결과가 다른 사건에 영향을 주지 않을 때 A와 B를 독립사건이라고 한다.

$$P(B|A) = P(B|A^c) = P(B)$$

두 사건 A와 B에서 한 사건의 결과가 다른 사건에 영향을 줄 때 A와 B를 종속사건이라고 함

$$P(B|A) \neq P(B|A^c) \neq P(B)$$

확률과 통계

➢ 조건부 확률

- 조건부 확률(conditional probability)은 어떤 사건 A가 일어났다는 조건하에 다른 사건 B가 일어날 확률

$$P(B \mid A)$$

- 사건 A가 일어났을 때, 사건 B에 대한 조건부 확률 법칙

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

- P(B|A): A 조건하에 B가 일어날 조건부 확률
- P (A, B) = P(AB) = P(A∩B): 함께/동시에 일어날 결합 확률(joint probability)
- P(A): 특정 사건 A에만 주목한 주변 확률(marginal probability)

확률과 통계

➢ 조건부 확률

- 예를 들어 ABC라는 고등학교에서 전체 학생의 65%가 안경을 썼고, 그 중 남학생이 안경을 쓴 경우는 45%임
- 이 학교의 학생 중 임의로 뽑은 1명이 안경을 쓴 학생이었을 때, 그 학생이 남학생일 확률을 구해 보자
- 전체 학생 중 안경을 쓴 학생(P(A))이 0.65고, 남학생이면서 안경을 쓴 학생(P(A∩B))이 0.45임
- 안경을 쓴 학생이면서 남학생일 확률은 다음과 같음

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.45}{0.65} = \frac{9}{13}$$

확률과 통계

➢ 조건부 확률

연습 문제

(1) 주머니 속에 흰색 공 네 개와 붉은색 공 여섯 개가 있습니다. 공을 한 개씩 두 번 꺼낼 때 다음 각 경우에서 두 개가 모두 흰색 공일 확률을 구하세요.

① 처음 꺼낸 공을 다시 넣지 않은 경우

② 처음 꺼낸 공을 다시 넣는 경우

확률과 통계

➢ 확률변수란

확률변수

- 실험 결과에 따라 표본 공간의 각 원소에 실수 값 하나를 대응시켜 주는 것을 **확률변수**(random variable)라고 한다.

X가 0일 때: 뒤뒤

X가 1일 때: 앞앞

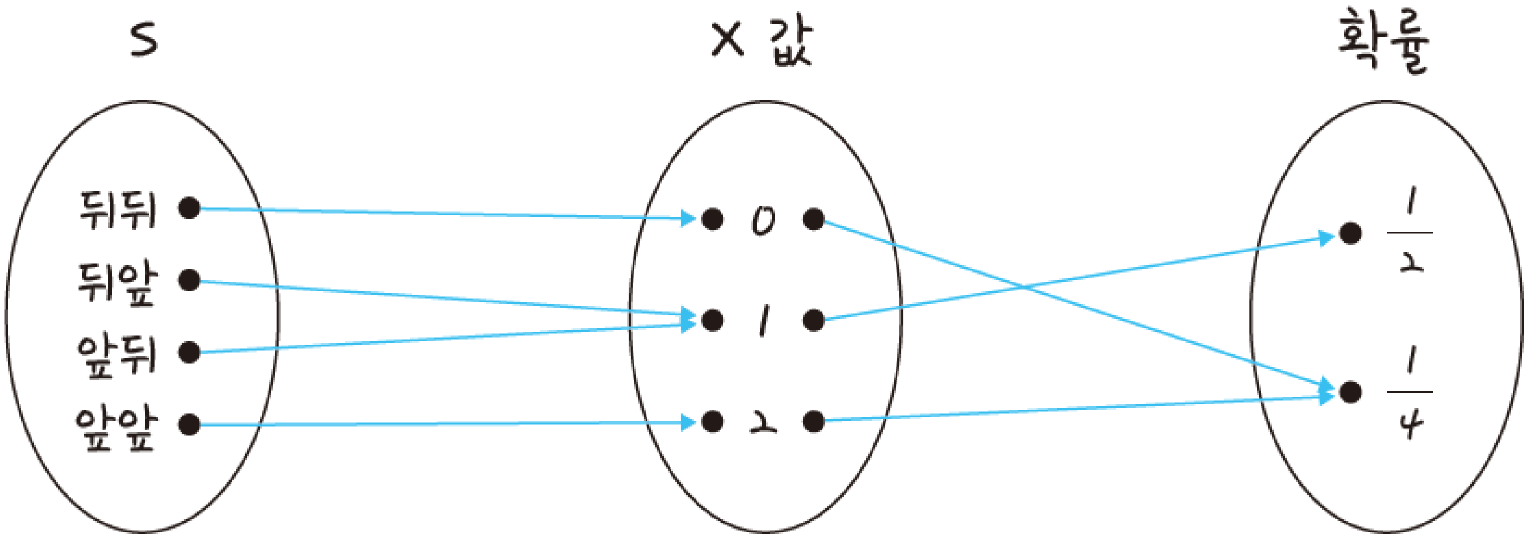
X가 2일 때: 앞뒤, 뒤앞

확률과 통계

➢ 확률변수란

● X가 0과 1일 수 있는 확률은 각각 $\frac{1}{4}$ 이고, X가 2일 수 있는 $\frac{2}{4}=\frac{1}{2}$ 임

▼ 그림 15-1 확률변수



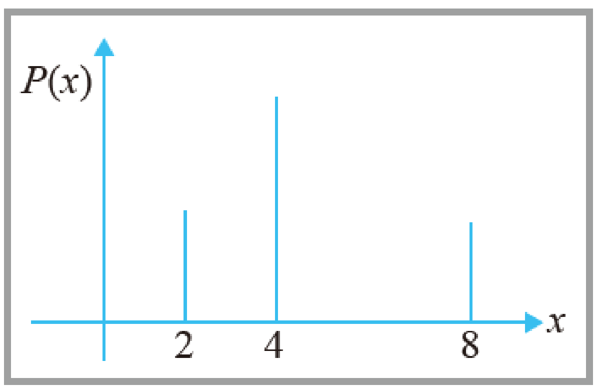
즉, 확률변수는 표본 공간의 모든 표본에 대해 어떤 실수 값을 할당하는 것으로 확률변수의 종류에는 다음과 같이 이산확률변수(discrete random variable)와 연속확률변수(continuous random variable)가 있다

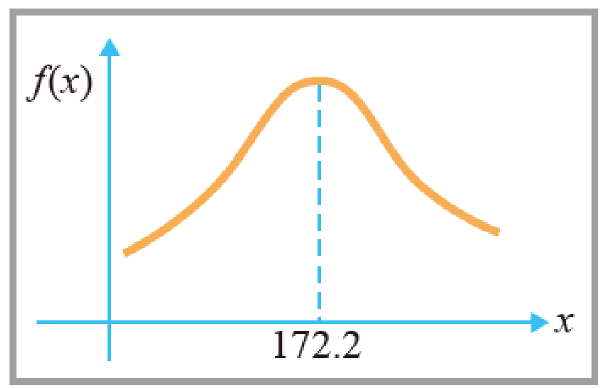
확률과 통계

확률변수

이산확률변수

연속확률변수





예를 들어 확률변수 X가 홍길동의 키를 나타낼 때 키가 178과 가깝다고 한다면, 홍길동의 키에 따른 확률은 다음과 같이 나타낼 수 있다.

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

확률과 통계

➢ 확률변수란

확률분포

- 확률분포(probability distribution)란 확률변수의 조합으로 생기는 확률 값의 분포를 그래프로 나타낸 것
- 예를 들어 동전을 두 번 던져 뒷면이 나오는 확률변수를 x라고 했을 때, 각 상태 공간 값이 나올 확률분포는 표1 5-1과 같음

표 15-1 확률변수 x의 상태 공간 값

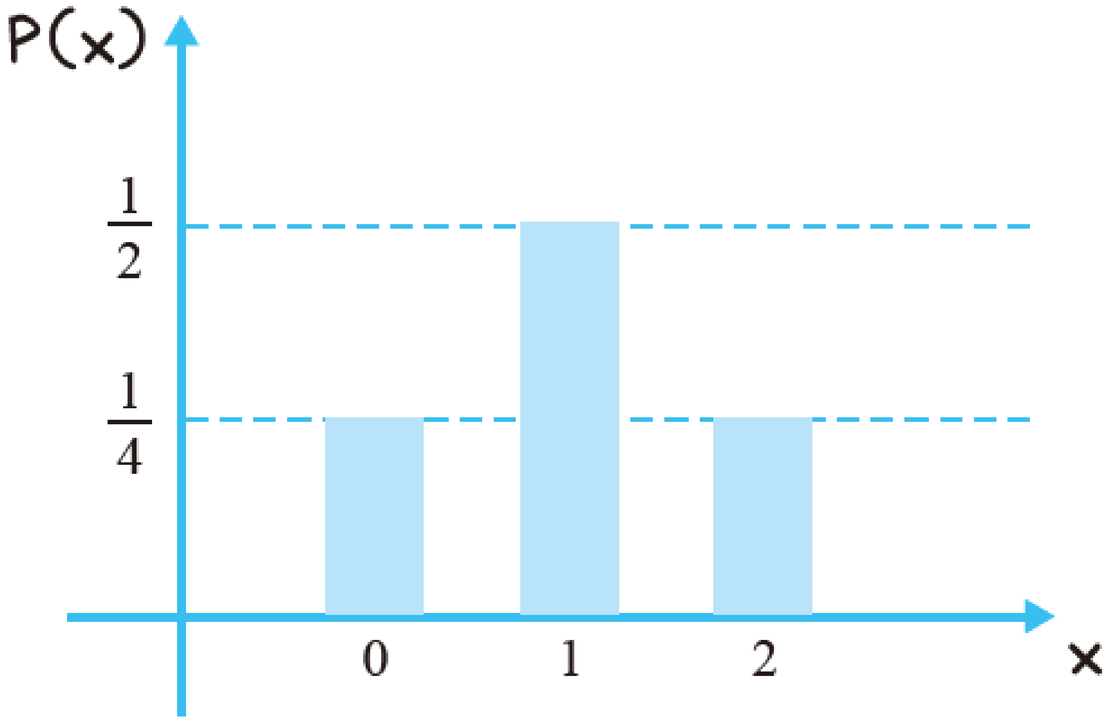
x	0	1	2
P(x)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

확률과 통계

➤ 확률변수란

- 동전을 두 번 던져 뒷면이 나오는 수를 x라고 했을 때, 확률분포의 그래프는 그림 15-3과 같음

그림 15-3 확률분포 그래프



확률과 통계

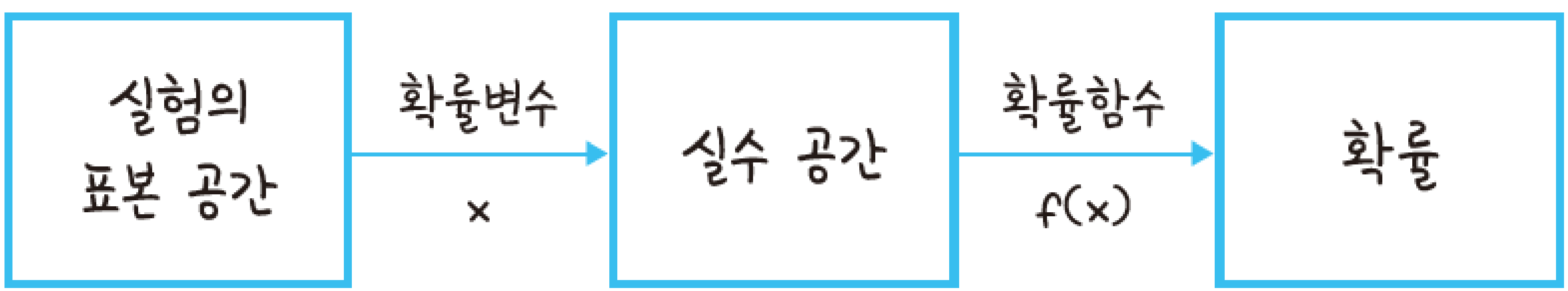
➤ 확률변수란

- 예를 들어 동전 한 개를 두 번 던지는 실험에서 앞면이 나오는 횟수를 확률변수 X라고 할 때, 확률변수 X가 취할 수 있는 값은 0, 1, 2고 이때 각 확률은 다음과 같음

(1) X가 0일 때 (뒤 $P(X=0) = \frac{1}{4}$

(2) X가 1일 때 (앞뒤, 뒤 $P(X=1) = \frac{2}{4} = \frac{1}{2}$

(3) X가 2일 때 (앞 $P(X=2) = \frac{1}{4}$

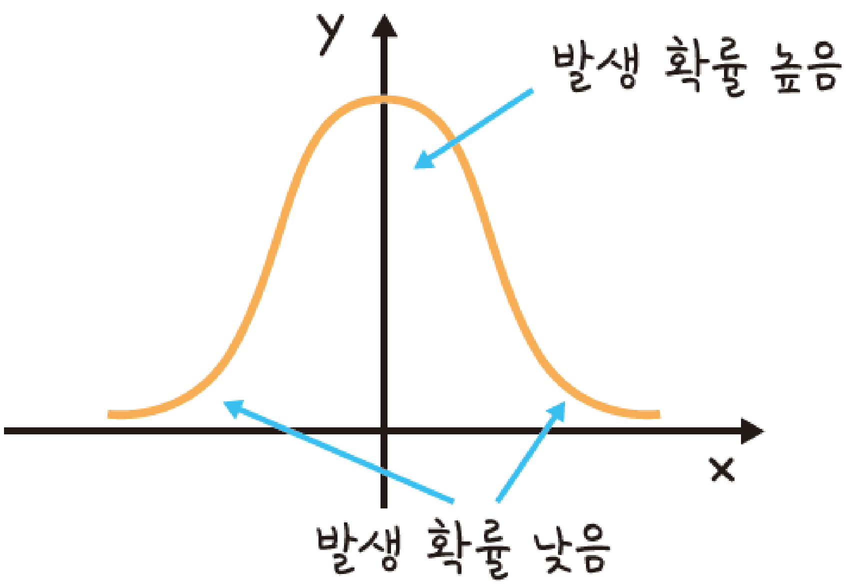


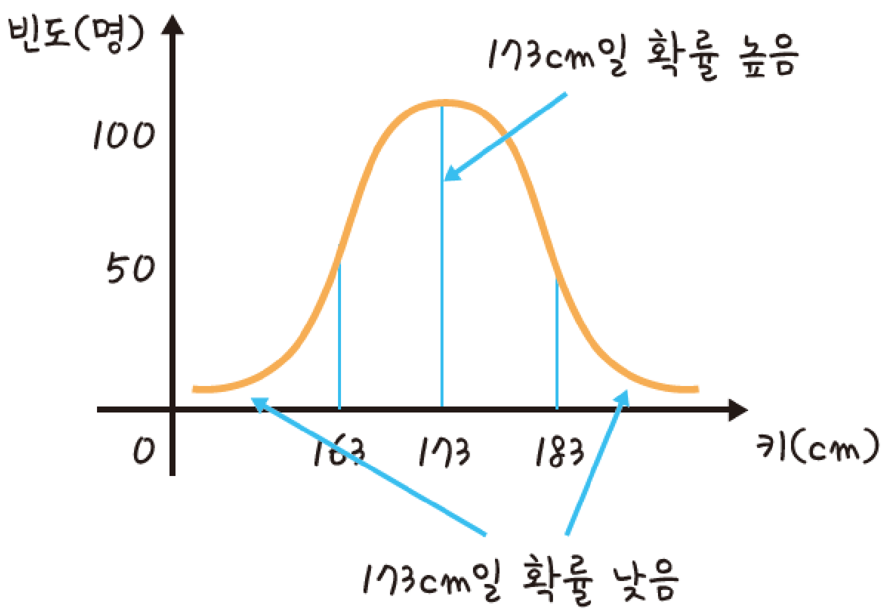
확률과 통계

➤ 확률분포의 유형

정규분포

- 평균에 가까울수록 발생할 확률이 높고, 평균에서 멀어질수록 발생할 확률이 낮게 나타나는 분포를 의미





확률과 통계

➢ 정규 분포의 시각화

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

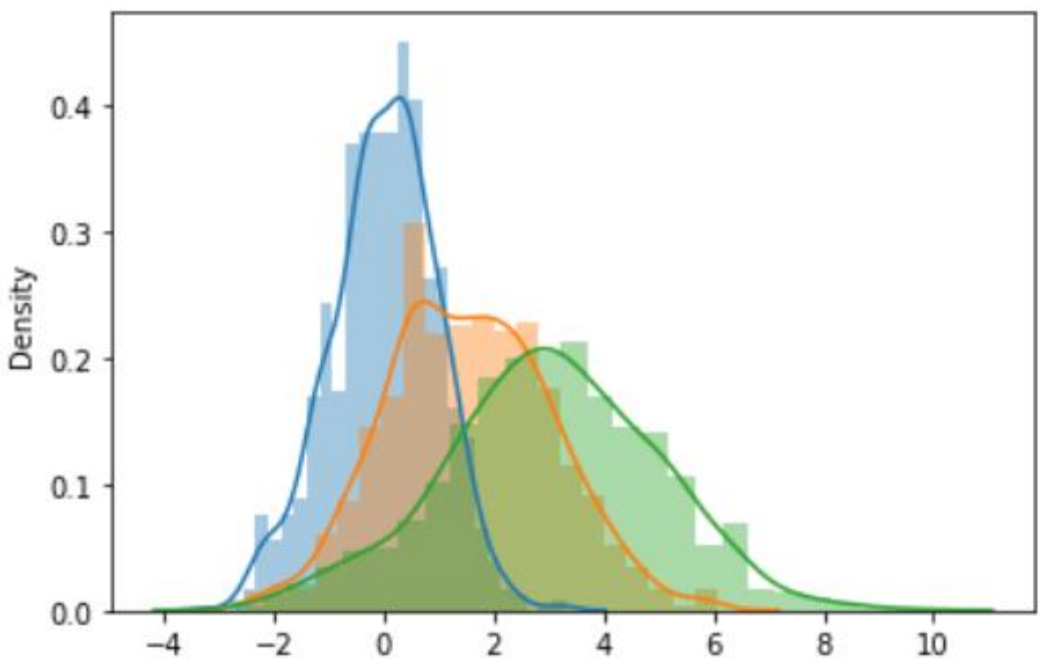
```
mu1, sigma1 = 0.0, 1.0
mu2, sigma2 = 1.5, 1.5
mu3, sigma3 = 3.0, 2.0
```

```
# random.normal() => 정규 분포로부터 얻은 임의의 샘플을 반환
y1 = np.random.normal(mu1, sigma1, 1000)
y2 = np.random.normal(mu2, sigma2, 1000)
y3 = np.random.normal(mu3, sigma3, 1000)
```

확률과 통계

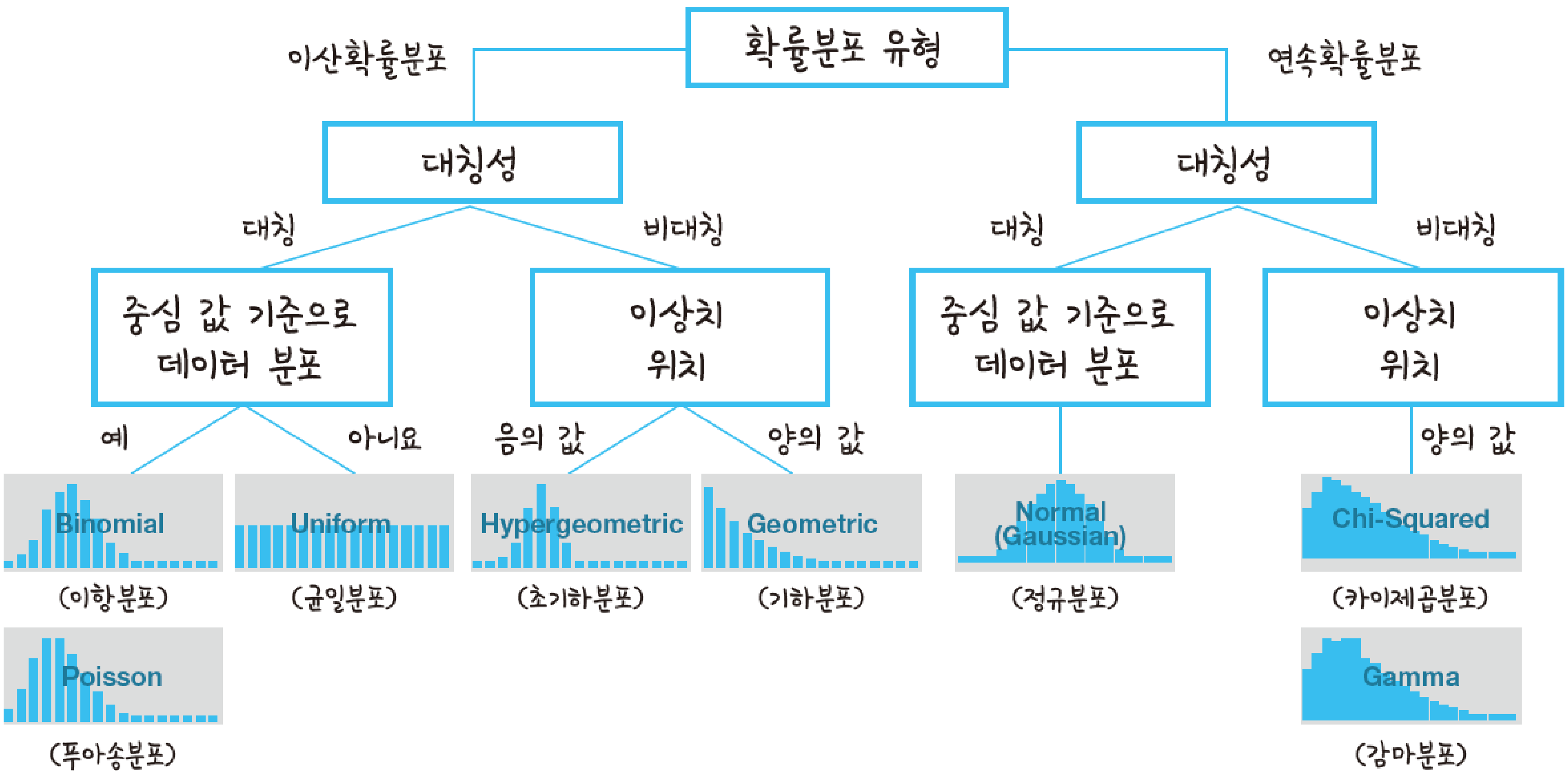
➢ **정규 분포의 시각화**

```
sns.distplot(y1);
sns.distplot(y2);
sns.distplot(y3);
```



확률과 통계

➢ 확률분포의 유형



확률과 통계

➢ 확률분포의 유형

표준정규분포(standard normal distribution)

- 정규분포의 평균을 '0'으로 만들고 표준편차를 '1'로 만들어서 표준화할 수 있다.
- 평균을 0으로 만들고 표준편차를 1로 만드는 방법은 간단함
- 다음과 같이 수집한 개별 데이터의 확률변수 X에서 그 데이터 전체의 평균(μ)을 빼고 표준편차(σ)로 나누면 된다.

표준화된 개별 데이터를 표준화 점수(Z-score)라고 하며, 표준화 점수는 평균이 0이고 표준편차가 1인 정규분포의 확률변수가 된다.

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

확률과 통계

대푯값

- 대푯값은 데이터를 가장 잘 설명하는 대표적인 값을 의미한다.
- 가장 자주 쓰는 대푯값으로 평균, 중앙값, 최빈값 등이 있다.

평균(mean : μ)

- 평균은 집단에서 중심 경향을 나타내는 수학적 척도로, 표본을 모두 더한 후 표본 개수로 나눈 값.

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

확률과 통계

중앙값(median)

- 중앙값은 주어진 값들을 크기대로 정렬했을 때 가장 중앙에 위치하는 값
- 전체 자료 개수가 홀수이면 중앙에 있는 값이 중앙값이고, 짝수이면 중앙에 있는 두 값의 평균을 중앙값으로 한다.

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 $\frac{(7+7)}{2} = 7$

최빈값(Mode)

- 최빈값은 가장 많이 관측되는 수, 즉 주어진 값 중에서 가장 자주 나오는 값
- 예를 들어 주사위를 열 번 굴려 1, 1, 2, 3, 3, 4, 6, 6, 6, 6이 나왔을 때, 가장 많이 등장한 '6'이 최빈값이 된다
- 최빈값의 장점은 가장 많이 발생하는 값을 구할 때 유용하다.

확률과 통계

파이썬 코드 구현

```
import numpy as np
import statistics as st
```

```
# 표본 생성
data = np.random.randint(1, 101, size=100)
data
```

```
array([81, 60, 55,  6, 89, 61, 10, 90, 25, 91,  7, 33, 92, 62, 13, 14, 21,
       34,  2, 32, 47, 44, 27, 25, 69, 90, 72,  8, 57, 83, 48, 39,  8, 74,
        7, 88, 51, 74, 94, 55, 18,  2, 76, 34, 72,  1, 68, 95, 88, 32, 23,
       64, 34, 85, 87, 35, 91, 58, 95, 50, 57, 14, 56,  7, 63, 42, 49, 74,
       27, 45,  9, 13, 76, 43, 27, 58, 47,  3, 46, 18, 99, 91, 45, 86, 27,
       22, 73, 66, 83, 53, 27, 31, 68, 78, 44, 44, 78, 17, 40, 91])
```

확률과 통계

파이썬 코드 구현

```
print(f' 평균 = { st.mean(data)}')
print(f' 중앙값 = { st.median(data)}')
print(f' 최빈값 = { st.mode(data)}')
```

```
평균 = 49
중앙값 = 48.5
최빈값 = 27
```

```
print(f' 최빈값 갯수 = {list(data).count(st.mode(data))}')
```

```
최빈값 갯수 = 5
```

확률과 통계

편차(Deviation)

: 변량-평균. 각 변량들이 평균으로부터 얼마나 떨어져 있는지 알려준다.

: 편차의 합은 0

분산(Variance)

: 편차의 제곱의 평균. 데이터의 퍼짐 정도

표준편차(Standard Deviation)

: 분산에 루트를 적용한다. 데이터의 퍼짐정도

확률과 통계

➢ 모집단분포와 확률표본

모집단과 모수

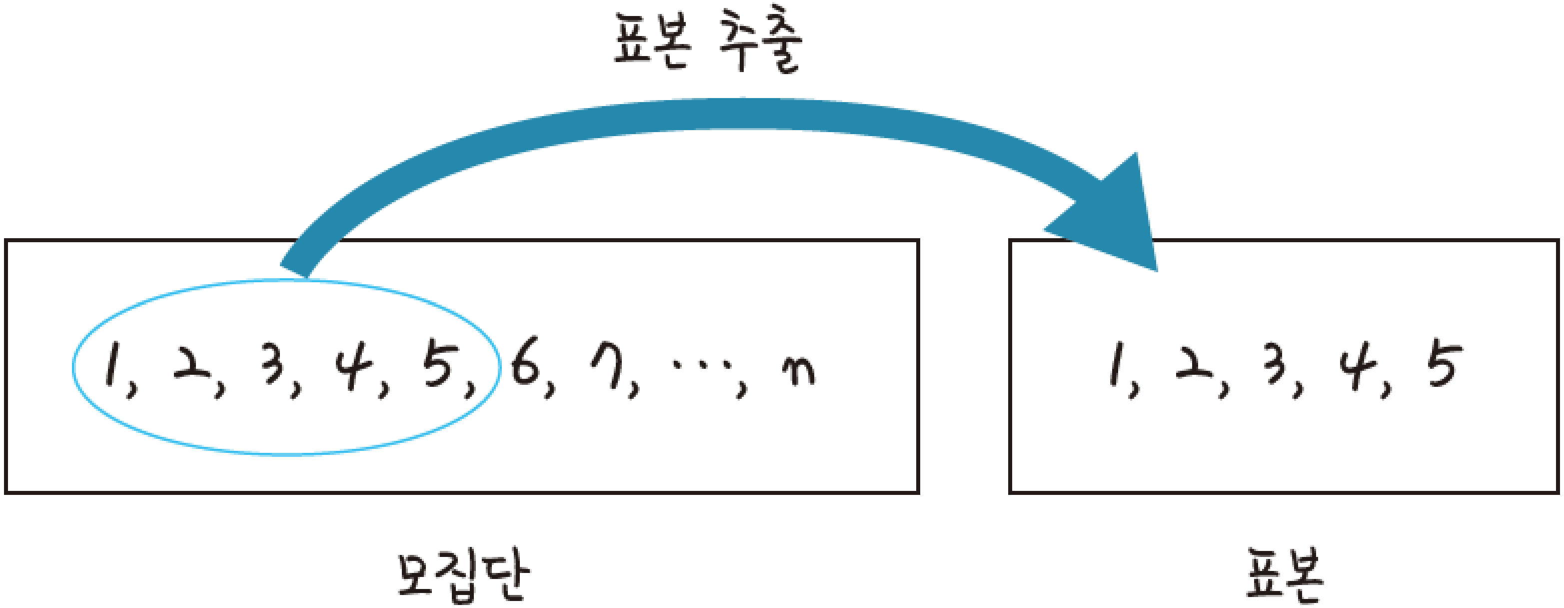
- 모집단은 어떤 통계적 실험의 대상이 되는 모든 대상물임
- 즉, 연구 대상(학생들의 수학 평균 점수를 조사하고 싶다)을 구성하는 모든 데이터(모든 학생의 수학 점수)라고 이해하면 됨
- 이때 모집단을 구성하는 데이터가 이루는 확률 분포가 **모집단분포**(population distribution)임

➢ 모집단분포와 확률표본

모집단과 확률표본

- 모집단의 모수(특성)를 알 수 있는 방법은 표본을 추출해서 모집단의 모수를 추정하는 것

그림 15-12 모집단과 표본



확률과 통계

➢ 회귀분석

- 독립변수와 종속변수는 인과 관계가 있다

그림 15-13 독립변수와 종속변수의 관계