파이썬 확률 통계

01 확률

목차 01 확률

1. 확률과 사건의 개념
2. 순열과 조합
3. 조건부 확률과 독립
4. 확률 분포

1. 확률과 사건의 개념

확률

여러 가능한 결과 중 하나 또는 일부가 일어날 가능성

0과 1사이의 값으로 정의

동전을 던졌을 때
앞면이 나올 가능성

=

동전을 던졌을 때
앞면이 나올 확률 0.5

확률의 용어

실험(Experiment) 또는 시행 (Trial)	여러 가능한 결과 중 하나가 일어나도록 하는 행위
표본공간 (Sample Space)	실험에서 나타날 수 있는 모든 결과들을 모아둔 집합 (Ω, S)
사건 (Event)	표본공간의 일부분(부분집합) 사건 A가 일어날 확률 : P(A) 또는 Pr(A)

동전을 던지는 실험

- 앞면 : H, 뒷면 : T, 표본공간 $\Omega = \{H, T\}$으로 표시
- 앞면이 나오는 사건은 $A = \{H\}$ 이므로 $P(A) = 0.5 = 1/2$

확률의 용어

추출 방법

복원추출

모든 시행에서
똑같은 상황으로 시행하는 방법

주머니에서 공을 꺼내 확인한 후
다시 넣고 다음 공 꺼내기

비복원추출

앞의 시행이
다음 시행에 영향을 주는 방법

주머니에서 공을 꺼내 확인한 후
다시 넣지 않고 다음 공 꺼내기

확률의 중요 구성요소

표본공간(Ω)

실험에서 나타낼 수 있는
모든 결과를 나열한 집합

사건(A)

모든 결과들 중 일부분
표본공간의 부분집합

확률(P)

모든 결과 중
사건이 발생하는 가능성을
0과 1사이의 값으로 정의

$$P(A) = \frac{\text{사건}A\text{의 원소의 수}}{\text{표본공간}(\Omega)\text{의 원소의 수}}$$

경우의 수

- 표본공간에서 사건 A가 발생할 확률
  - $P(A) = \frac{A\text{에 속하는 결과의 수}}{\text{총 가능한 결과의 수}}$
- 사건 A의 확률을 정의하기 위해 A에 속하는 결과의 수 파악 필요
- 사건의 원소의 개수(사건에 속하는 결과의 수)

  = 1회 시행에서 일어날 수 있는 사건의 가짓수

  = 사건의 경우의 수

사건의 기본적인 연산

- $\boldsymbol{A}$의 여사건
  - 사건 A에 포함되지 않은 사건들의 집합
  - A^c 로 표시
- $\boldsymbol{A}$와 $\boldsymbol{B}$의 합사건
  - 사건 A 혹은 B에 포함되는 사건들의 집합
  - $A \cup B$로 표시
- A와 B의 곱사건
  - 사건 A와 B에 동시에 포함되는 사건들의 집합
  - $A \cap B$로 표시
- 배반사건
  - 동시에 일어날 수 없는 두 사건
  - $A \cap B = O$인 두 사건

경우의 수의 계산

- 합의 법칙
  - 두 사건 A와 B가 일어나는 경우의 수가 각각 m과 n
  - 두 사건 A 와 B 가 동시에 일어나지 않음
  - 사건 A 또는 B 가 일어나는 경우의 수는 $m + n$
- 곱의 법칙
  - 두 사건 A 와 B 가 일어나는 경우의 수가 각각 m 과 n
  - 두 사건 A 와 B 가 동시에 또는 잇달아 일어남
  - 이때 경우의 수는 $m \times n$

팩토리얼(!)

- 1부터 어떤 양의 정수 n까지의 정수를 **모두 곱한 것**
  - $0! = 1$
  - $1! = 1$
  - $n! = n \times (n-1)!$

```
def fac(n):
    if n == 0:
        return 1
    if n == 1 :
        return 1
    else:
        return n * fac(n-1)

# if 문을 활용하여 n 이 0혹은 1
인 경우, 1을 반환하며 그 외의
경우는 n * (n-1)!를 반환
```

확률의 공리

- 공리 : 증명을 필요로 하지 않거나 증명할 수 없지만
  직관적으로 자명한 진리의 명제인 동시에 다른 명제들의 전제가 되는 명제

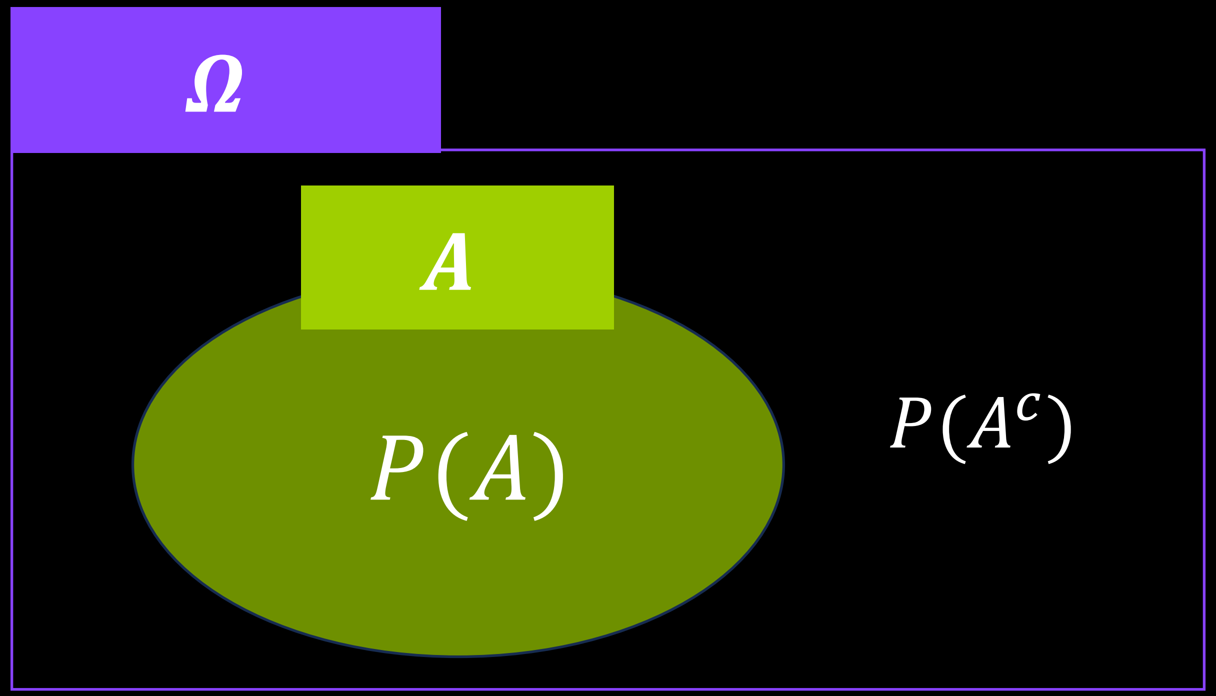
- 확률의 공리

1. 모든 사건 A에 대하여 $0 \leq P(A) \leq 1$
   - 어떤 확률도 0보다 작거나 1보다 클 수 없음
2. 표본공간 Ω 에 대하여 $P(\Omega) = 1$
   - 전체 표본공간, 즉 모든 확률의 합은 1임
3. 사건 $A_1, A_2, \ldots$이 서로 배반사건일 때 $P(U_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
   - 각 사건들의 교집합은 공집합이므로 서로 배반인 사건들이 일어날 전체 확률은 각각의 확률을 더한 것과 같음

확률의 정리

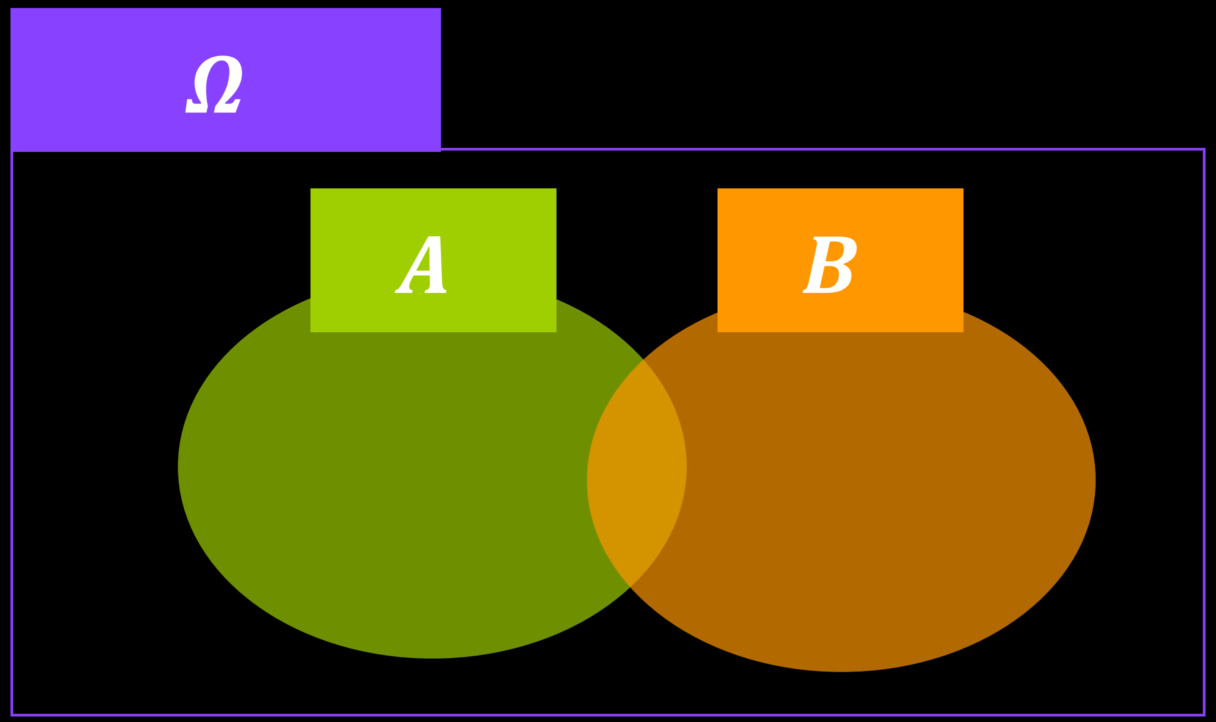
1. 사건 A의 여집합 A^c 에 대해

$$P(A^c) = 1 - P(A)$$

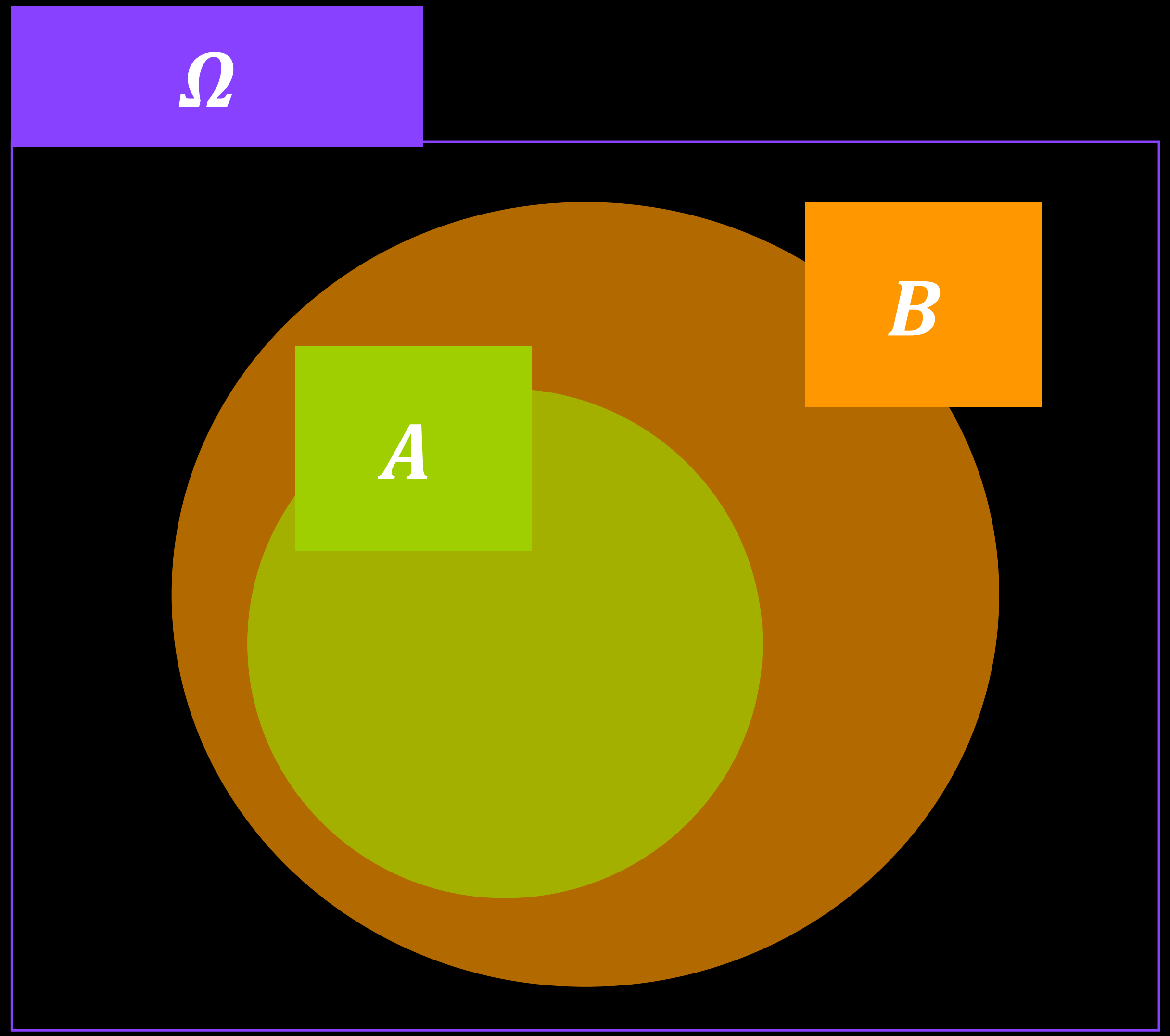


2. 임의의 두 사건 A와 B에 대하여

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



3. $A \subset B$ 일 때 $P(A) \leq P(B)$



2. 순열과 조합

순열

- 곱의 법칙에 의해 총 가능한 경우의 수

  = n개의 서로 다른 원소 중 k개를 선택하여 배열하는 경우의 수

  = 순열

- $n(n-1)\cdots(n-k+1) = {}_{n}P_{k}$

순열

- [] 안에 서로 다른 n개의 원소를 주고

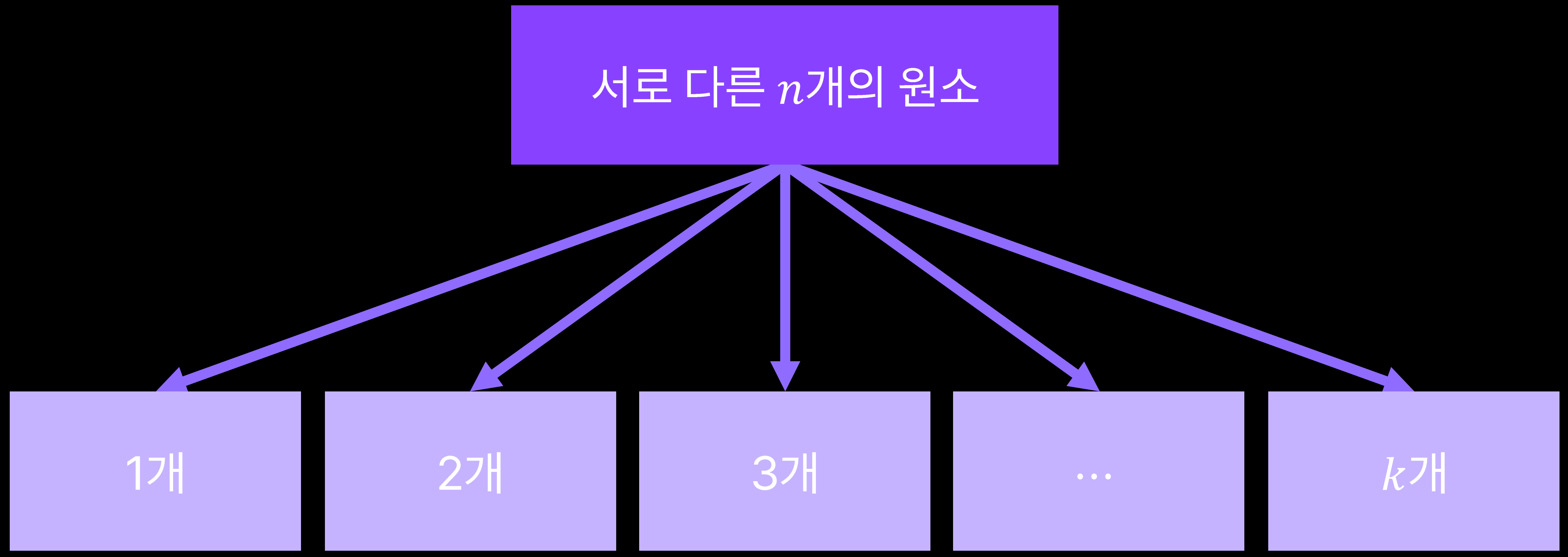
  이 원소들 중 k개를 순서를 고려하여 뽑는 경우의 수 계산

```
from itertools import permutations

list(permutations([n], k))
```

조합

- 서로 다른 n개의 원소에서 k개를 **순서에 상관없이** 선택하는 방법
- $\dfrac{(n\text{개의 서로 다른 원소 중 } k\text{개를 선택하여 배열하는 경우의 수})}{(k\text{개의 원소를 나열하는 경우의 수})} = \dfrac{n(n-1)\cdots(n-k+1)}{k!} = {}_nC_k$



조합

- [] 안에 서로 다른 n개의 원소를 주고

  이 원소들 중 k개를 순서를 고려하지 않고 뽑는 경우의 수 계산

```
from itertools import combinations

list(combinations([n], k))
```

순열과 조합의 차이

순열	조합
순서가 있음	순서가 없음
반장/부반장 뽑기	2명의 대표 선발

중복순열

- 서로 다른 n개의 원소 중에서 중복을 허용하여 r개를 뽑아 일렬로 배열하는 경우
  - ${}_n\Pi_r = n^r$
- [] 안에 서로 다른 n개의 원소를 주고

  이 원소들 중 k개를 중복을 허용하면서 순서를 고려하여 뽑는 경우의 수 계산

```
from itertools import product
list(product([n], repeat = k))
```

중복조합

- 서로 다른 n개의 원소 중에서 중복을 허용하여 r개를 순서를 고려하지 않고 뽑는 경우
  - ${}_{n}H_{r} = {}_{n+r-1}C_{r}$
- [] 안에 서로 다른 n개의 원소를 주고

  이 원소들 중 k개를 **중복을 허용**하면서 **순서를 고려하지 않고** 뽑는 경우의 수 계산

```python
from itertools import combinations

list(combinations_with_replacement([n], k))
```

3. 조건부 확률과 독립

조건부 확률

- 특정한 사건의 확률을 구할 때, 다른 사건에 대한 정보가 주어지는 경우
- 다른 사건에 대한 정보를 이용하여 확률을 구하므로 기존 확률과 달라질 수 있음
- $P(B) \neq 0$인 사건 B에 대한 정보가 주어졌을 때 A가 발생할 조건부 확률을 $P(A|B)$라 하면

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

- $P(B|A)$ 와 $P(A)$를 이용해 사건 $A \cap B$의 확률 계산

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

조건부 확률 예시

- 수강생 30명을 성별과 응답(A, B) 에 따라 구분
- 이 반에서 임의로 학생 한 명을 선택하는 경우

응답\성별	남	여	계
A	10	5	15
B	8	7	15
계	18	12	30

- 선택된 학생의 응답이 A일 확률
  - 임의로 선택된 학생이 A 응답일 확률은 전체 학생이 30명, A 응답 학생이 15명이므로 15/30

$$\text{확률} = \frac{\text{구하려는 사건}}{\text{전체 경우의 수}}$$

- 여학생을 선택했을 때 응답이 A일 확률
  - 여학생의 수는 12명, A 응답 여학생은 5명이므로 5/12

$$\text{조건부확률} = \frac{\text{구하려는 사건이 일어날 확률}}{\text{정보가 주어진 사건의 확률}}$$

독립

- 일반적으로 두 사건은 서로 연관성이 있는 경우가 많음
  - 동전을 두 번 던지는 실험에서 A는 앞면이 2번 나오는 사건이라 하고
    B는 앞면이 1번 이상 나오는 사건이라고 할 때,
    사건 A가 발생하였다는 정보가 주어지면 사건 B는 반드시 발생했다고 할수있음
- 조건부 확률은 두 사건의 **연관성**에 따라 달라짐

사건 A와 사건 B 사이에 연관성이 없음	=	사건 B는 A에 아무런 영향을 주지 못함	=	$P(A\|B) = P(A)$
사건 A와 사건 B 사이에 어떤 연관성이 있음	=	사건 B는 A에 영향을 줄 수도 있음	=	$P(A\|B) \neq P(A)$

독립

- 두 사건 A와 B가 서로 독립일 때 사건 B가 A의 확률에 영향을 주지 않음

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

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

4. 확률 분포

확률변수

- 각각의 근원사건에 실수 값을 대응시킨 함수 ($X, Y, \cdots$ 처럼 대문자로 표시)
- 시행을 하기 전엔 어떤 값을 갖게 될 지 알 수 없다는 불확실성을 표현
- 확률변수가 가질 수 있는 값들이 무엇이며, 그 값들을 가질 가능성 또는 확률이 어떻게 분포되어 있는지를 0 이상의 실수로 나타낸 것
- 예시 : 동전을 2번 던졌을 때 앞면의 수를 나타내는 확률변수 X

$$P(X = 2) = P(\{\text{앞앞}\}) = \frac{1}{4}$$

$$P(X = 1) = P(\{\text{앞뒤}, \text{뒤앞}\}) = \frac{1}{2}$$

이산, 연속 확률변수

- 이산 확률변수
  - 확률변수의 값의 개수를 셀 수 있는 경우
- 연속 확률변수
  - 확률변수의 값이 연속적인 구간에 속하는 경우

이산 확률변수

- 확률변수의 값의 개수를 셀 수 있는 경우

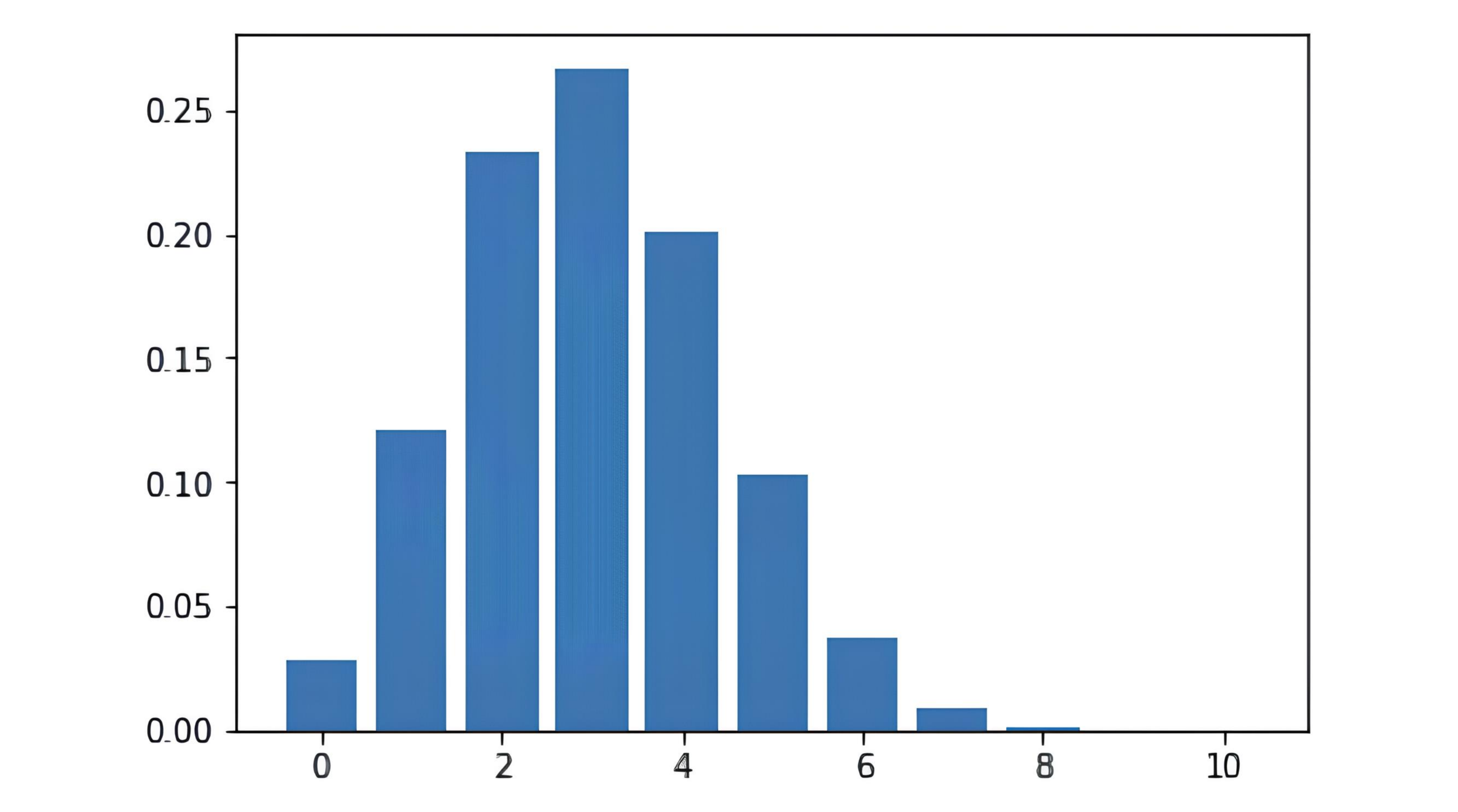
변수 x	확률 $f(x)$
x_1	$f(x_1)$
x_2	$f(x_2)$
$\cdots$	$\cdots$
x_k	$f(x_k)$
합계	1

확률 질량 함수

- 어떤 확률변수 x가 갖는 확률을 나타내는 함수
  - $y = f(x)$: x가 갖는 확률은 y이다
- 확률 질량 함수의 조건
  - 모든 x_i값에 대해 $0 \le f(x_i) \le 1$이고 $\sum_{all\ x_i} f(x_i) = 1$

이산 확률분포

- 이산 확률분포의 종류
  - 베르누이 분포, 이항분포, 기하분포, 음이항분포, 초기하분포, 포아송분포, …

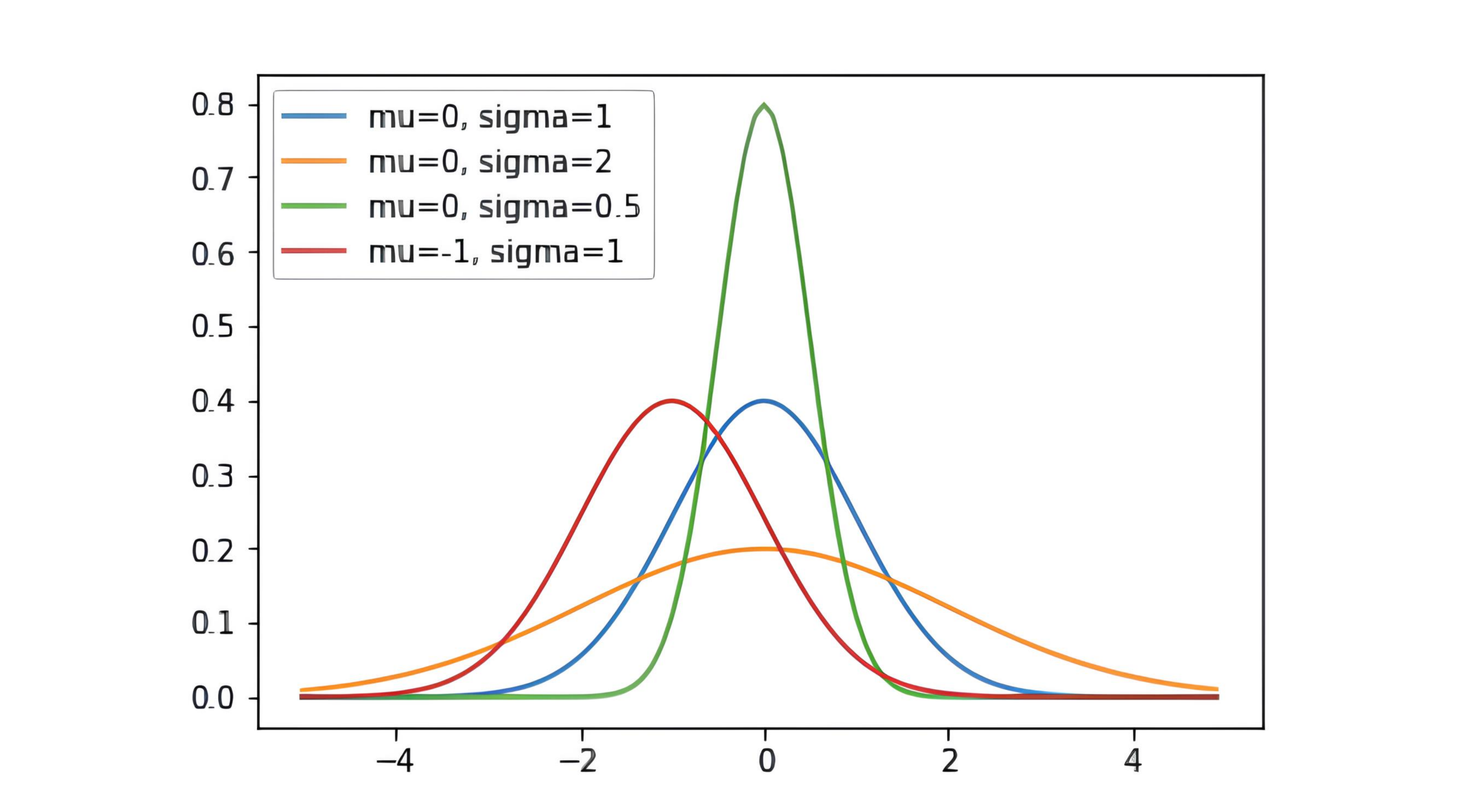


확률 밀도 함수

- 어떤 확률변수 X가 갖는 확률의 분포를 표현
- 어느 구간의 확률이 더 크고 작은 지 나타낼 수 있는 함수를 이용
- 확률 밀도 함수의 조건
  - 모든 x값에 대해 $f(x) \geq 0$: 모든 x값에 대해 확률 밀도 함수 값은 0보다 크거나 같다.
  - $P(a \leq X \leq b) = \int_a^b f(x)\,dx$: a~b까지 구간의 확률은 그 구간만큼 $f(x)$에서 적분한 값과 같다.
  - $P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1$: 전체 구간을 적분했을 때 확률 밀도 함수 값은 1이다.

연속 확률분포

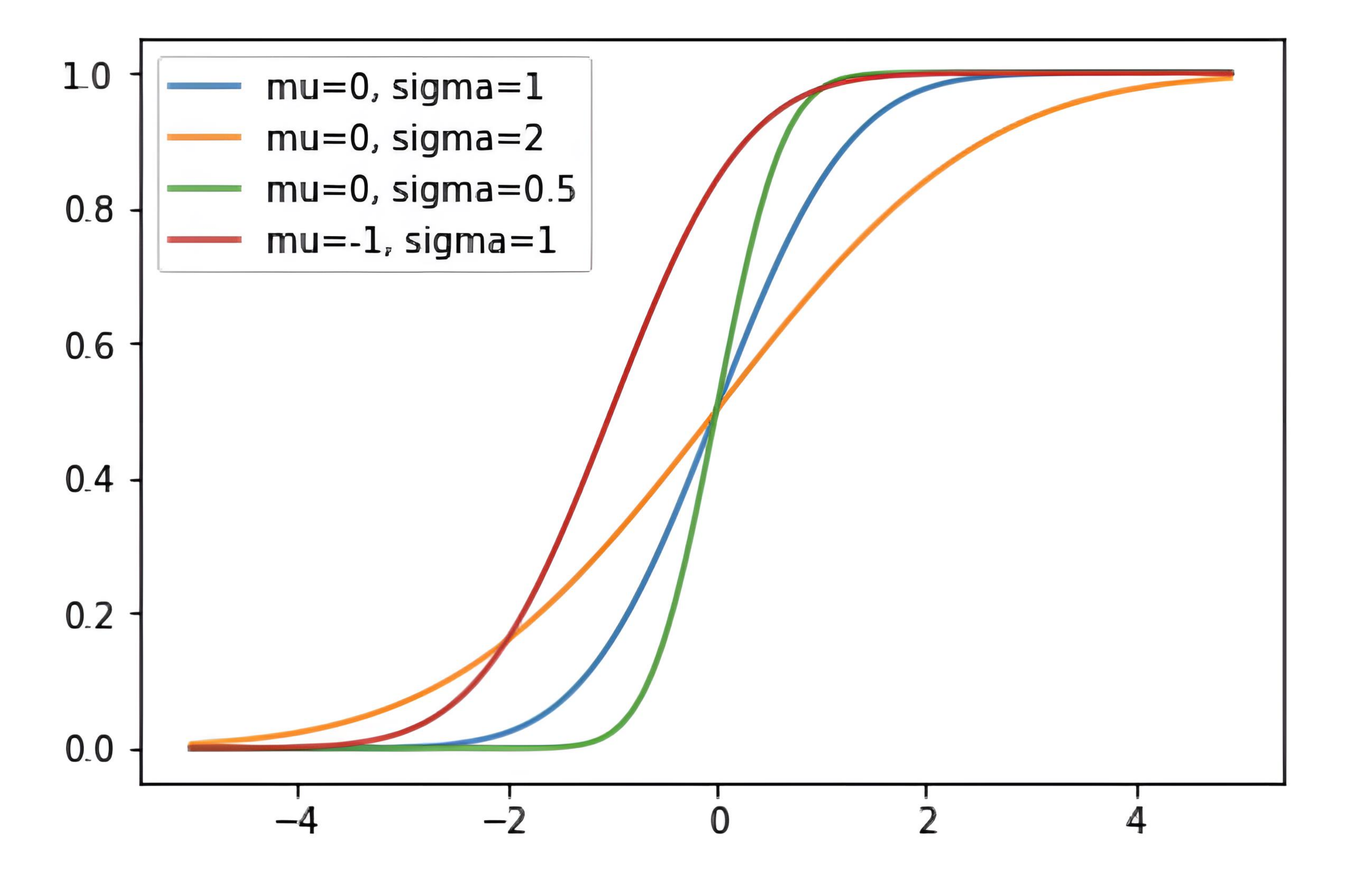
- 연속 확률분포 종류
  - 균일분포, 지수분포, 감마분포, 정규분포, 베타분포, …



누적분포함수

- X가 가질 수 있는 가장 작은 값부터 x까지 해당하는 확률질량함수의 값을 누적해서 더한 것
  - $F(x) = P(X \leq x)$
  - 이산 확률변수의 누적분포함수 : $F(x) = \sum_{y \leq x, y \in X(s)} p(y)$
  - 연속 확률변수의 누적분포함수 : $F(x) = \int_{-\infty}^{x} f(y)\, dy$

연속형 누적분포함수


mu=0, sigma=1
mu=0, sigma=2
mu=0, sigma=0.5
mu=-1, sigma=1
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4

이산형 누적분포함수