

이공학도를 위한 확률 및 통계학

-3장 확률변수와 확률분포-

서진우(jinwoo@pel.sejong.ac.kr)

세종대학교 프로토콜공학연구실

목 차

- 보충
- 확률변수의 개념
- 확률분포와 누적분포함수
- 결합확률분포

목 차

- 보충
- 확률변수의 개념
- 확률분포와 누적분포함수
- 결합확률분포

보충

- 누적분포함수(CDF, Cumulative Distribution Function) (1/6)
 - 정의
 - 실수 x 에 대해 확률변수 X 가 x 이하일 확률을 나타내는 함수
 - e.g., 로그인 실패 횟수가 5회 이하일 누적 확률
 - 특징
 - 비감소 함수, 주어진 확률변수 뿐만 아니라 모든 실수에서 정의
 - 이산형 확률분포에서의 누적분포함수 $F(x)$
 - $F(x) = P(X \leq x) = \sum_{t \leq x} f(t), -\infty < x < \infty$
 - 연속형 확률분포에서의 누적분포함수 $F(x)$
 - $F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt, -\infty < x < \infty$
 - 필요성
 - 연산의 단순화
 - 프로그래밍/알고리즘 표준화

보충

- 확률분포 시각화 (1/5)

- 확률질량함수도

- 정의

- 확률 값을 x 축과 점선이나 실선으로 연결하여 이산형 확률 분포를 나타낸 그림

- 장점

- 개별 사건의 정확한 확률을 직관적으로 확인 가능

- 단점

- 연속적인 값에는 적용 불가

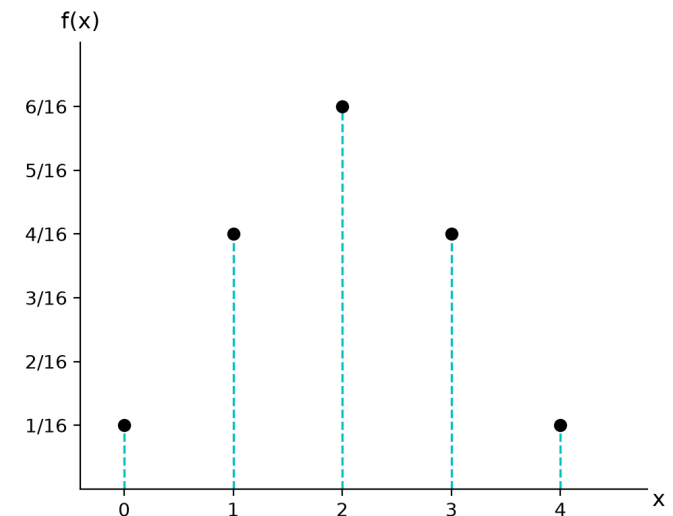
- 활용 예시

- 무차별 대입 공격 탐지

- e.g., 특정 IP가 1분 동안 로그인 실패를 5회, 10회, 15회 할 확률 계산

- 디지털 포렌식

- e.g., 악성코드 파일 내에서 특정 바이트가 나타나는 빈도 분석



<그림 1>
확률질량함수도

보충

- 확률분포 그래프 (2/5)

- 확률히스토그램

- 정의

- 확률질량함수도에서 점 대신 직사각형을 사용해 나타낸 그림

- 장점

- 데이터의 밀집도 파악 가능

- 단점

- 구간 설정에 따라 그래프의 모양이 왜곡될 수 있음

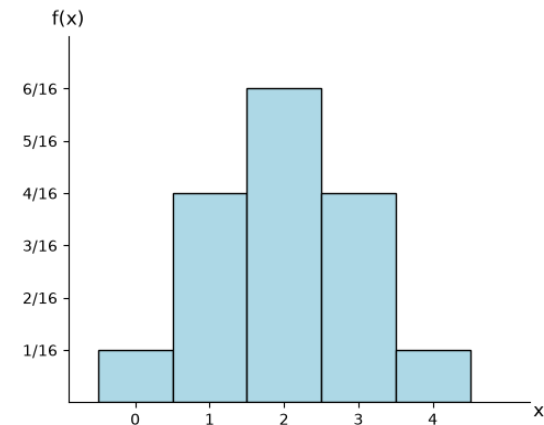
- 활용 예시

- 트래픽 베이스라인 모니터링

- e.g., 사내 망의 시간대별 패킷 크기 분포를 시각화하여, 평소와 다른 기형적인 모양이 나타나는지 감시

- 보안 장비 리소스 모니터링

- e.g., 방화벽 CPU 사용량 분포 시각화



<그림 2>
확률히스토그램

보충

- 확률분포 그래프 (3/5)

- 확률 밀도 함수

- 정의

- 특정 지점의 y 축 값은 확률이 아닌 밀도이며, 특정 구간의 면적이 확률인 그림

- 장점

- 데이터가 밀집된 구간과 희소한 구간을 매끄러운 곡선으로 표현 가능

- 단점

- 확률 계산 시, 적분을 해야 하므로 번거로움 발생

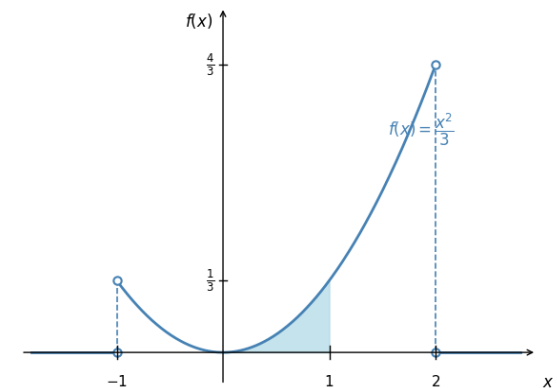
- 활용 예시

- 사용자 및 개체 행동 분석

- e.g., 특정 임직원의 평소 시스템 접속 시간 분포 곡선 생성 후, 해당 이상 징후 탐지

- APT 공격 탐지

- e.g., C&C 서버와 통신하는 주기의 밀도 분석



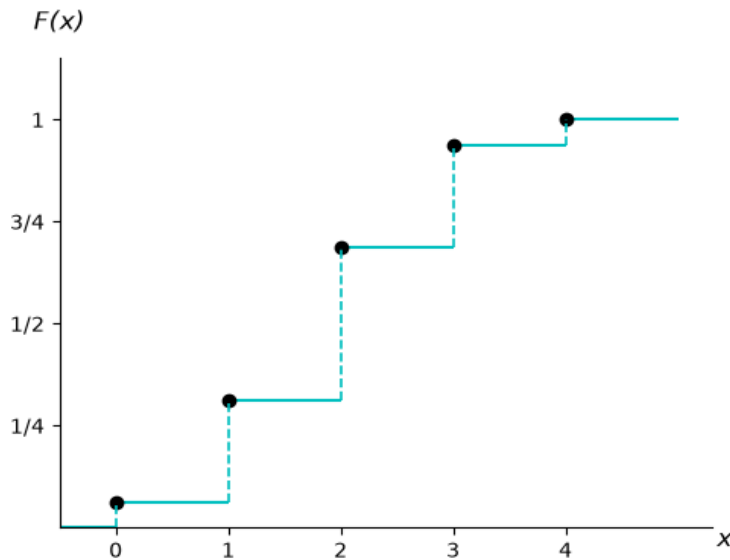
<그림 5>
확률밀도함수

보충

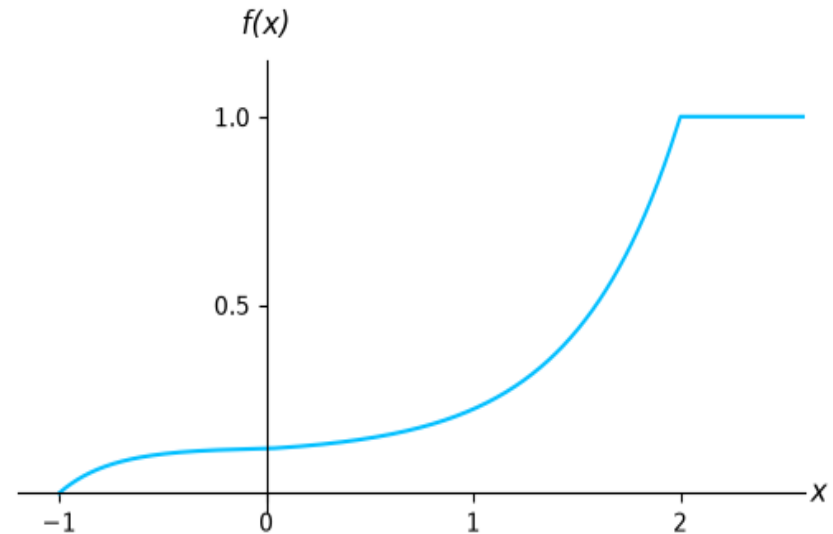
- 확률분포 그래프 (4/5)
- 누적분포함수 (1/2)
 - 정의
 - 실수 x 에 대해 확률변수 X 가 x 이하일 확률을 나타내는 함수
 - 장점
 - 적분 없이 단순 뺄셈을 통해 구간의 확률 계산 가능
 - 단점
 - 확률이 누적되어 특정 구간의 분포를 직관적으로 파악하기 어려움
 - 활용 예시
 - 리스크 관리 및 가설 검정
 - e.g., 위험 평가 시 “자산이 랜섬웨어에 감염되어 발생할 손실액이 1억 원 이하일 경우”의 누적 확률 계산
 - 실시간 탐지 알람 제어
 - e.g., 보안 솔루션이 탐지한 위협 점수를 누적분포함수에 대입하여 “현재 이벤트가 상위 0.1%에 해당하는 극단적 위협”인지 즉시 판정하여 차단 알림 전송

보충

- 확률분포 그래프 (5/5)
- 누적분포함수 (2/2)



<그림 6>
이산형 누적분포



<그림 7>
연속형 누적분포

목 차

- 보충
- 확률변수의 개념
- 확률분포와 누적분포함수
- 결합확률분포

확률변수의 개념

- 확률변수(Random Variable) (1/9)

- 정의

- 표본공간 내 각 원소에 하나의 실수 값을 대응시키는 함수

- 특징

- 주로 대문자 표기하며, 그에 대응하는 하나의 값은 소문자 표기

- e.g., X : 주사위를 한 번 던져 나온 눈의 수, x : 실제 나온 값

- 확률변수가 특정 값을 가지는 사건은 표본공간의 부분집합과 동일

- e.g., 로그인 시도 100건(표본공간) 중 X ="악성 여부"일 때, 사건 $\{X = 1\}$ 은 실제 악성으로 판정된 시도들의 부분집합에 대응

- 종류

- 이산형 확률변수(Discrete Random Variable)

- 가변수

- 연속형 확률변수(Continuous Random Variable)

확률변수의 개념

- 확률변수(Random Variable) (2/9)

- 예제 3.1

4개의 붉은 공(R) 과 3개의 검은 공(B)이 들어 있는 항아리에서 연속적으로 2개의 공을 비복원 추출하는 실험에서 Y 를 붉은 공의 개수라 할 때, 출현 가능한 결과와 확률변수 Y 의 값 y 를 구하라.

- 풀이

- $Y = \{0, 1, 2\}$

표본공간	y
BB	2
BR	1
RB	1
RR	0

확률변수의 개념

- 확률변수(Random Variable) (3/9)

- 예제 3.2

공구보관소의 직원이 3명의 공장 종업원들에게 안전 헬멧을 임의로 꺼내주었다. 만일 스미스(S), 존스(J), 그리고 브라운(B)의 순서로 헬멧을 받을 때 헬멧을 받는 가능한 순서들을 나열하라. 그리고 M 을 헬멧이 원래 주인에게 지급되는 경우의 수라고 할 때, 확률변수 M 의 값 m 을 구하라.

- 풀이

- $M = \{0, 1, 3\}$

표본공간	m
SJB	3
SBJ	1
JSB	1
JBS	0
BSJ	0
BJS	1

확률변수의 개념

- 확률변수(Random Variable) (4/9)

- 가변수(Dummy Variable)

*베르누이 확률변수(Bernoulli Random Variable)
두 개의 가능한 값을 0과 1로 표현하는 확률변수.

- 정의

- 성별이나 지역처럼 숫자가 아닌 성질을 가진 데이터를 분석하기 위해 0과 1로 바꾸어 표현한 변수

- 특징

- 확률변수가 범주형인 경우 사용
 - 표본공간이 무한 개의 원소를 가질 때 사용

- 예제 3.3

공장에서 생산되는 부품들을 불량이나 양품으로 판정한다고 할 때, 베르누이 확률변수를 통해 확률변수 X 를 정의하라.

- 풀이

- $$X = \begin{cases} 1, & \text{부품이 불량일 경우} \\ 0, & \text{부품이 양품일 경우} \end{cases}$$

확률변수의 개념

- 확률변수(Random Variable) (5/9)

*가산 무한(Countable Infinite)
순서를 매겨 나열할 수 있는 상태.

- 이산표본공간(Discrete Sample Space)

- 정의

- 유한개 혹은 가산 무한개의 원소로 이루어진 표본공간

- e.g., 포트 스캔 시 열려있는 것으로 확인된 포트 번호의 집합
 $S = \{22, 80, 443, 3306, 8080\}$

- e.g., 비밀번호 무차별 대입 공격 성공 시점의 누적 시도 횟수 $S = \{0, 1, 2, 3, \dots\}$

- 이산형 확률변수(Discrete Random Variable)

- 정의

- 확률변수의 집합이 셀 수 있는 집합일 때의 확률변수

- e.g., 하루 동안 특정 서버에 대해 발생하는 로그인 실패 횟수 $X = \{0, 1, 2, 3, \dots\}$

- 특징

- 주로 계수자료(Count Data)를 나타낼 때 사용

- e.g., 하루 동안 방화벽에서 차단된 패킷 수

확률변수의 개념

- 확률변수(Random Variable) (6/9)

- 예제 3.4

로트(lot)의 합격 또는 불합격 판정을 위해 샘플링검사법(Sampling Plan)을 사용한다. 12개의 불량품이 있는 100개의 제품에서 10개를 독립적으로 추출하는 샘플링검사법을 생각해보자.

- 풀이
 - 10개 제품 표본에서 발견되는 불량품의 수를 확률변수 X
 - $X = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

확률변수의 개념

- 확률변수(Random Variable) (7/9)

- 예제 3.5

하나의 불량품이 발견될 때까지 공정으로부터 표본을 추출하는 샘플링검사법이 있다고 하자. 공정의 능력은 얼마나 많은 제품이 불량 없이 연속적으로 추출되는가에 달려있다. 불량품이 발견될 때까지 추출한 제품의 수를 나타내는 확률변수를 X 라고 하자. 양품을 N , 불량품을 D 로 나타낼 때, X 값이 1, 2, 3일 때의 표본공간(S)를 쓰시오.

- 풀이

- $X = 1$ 일 때, $S = \{D\}$
- $X = 2$ 일 때, $S = \{ND\}$
- $X = 3$ 일 때, $S = \{NND\}$

확률변수의 개념

- 확률변수(Random Variable) (8/9)
- 연속표본공간(Continuous Sample Space)
 - 정의
 - 실선의 특정 구간에 속한 모든 실수를 원소로 갖는 표본공간
 - e.g., 웹 서버의 특정 API 응답 지연 시간의 범위 $S = \{t | 0 \leq t \leq 10\}$
- 연속형 확률변수(Continuous Random Variable)
 - 정의
 - 확률변수가 연속적인 구간 내의 값을 취할 때의 확률변수
 - e.g., 악성코드 탐지 엔진이 의심 파일을 정밀 스캔하여 최종 분석을 완료할 때까지 소요되는 시간
 - 특징
 - 주로 측정자료(Measured Data)를 나타낼 때 사용
 - e.g., 높이, 무게, 온도, 거리, 시간

확률변수의 개념

- 확률변수(Random Variable) (9/9)

- 예제 3.6

두 상황에 따른 확률 변수 X 의 값 x 의 범위를 구하라.

1. 통신판매 광고에 반응하는 사람들의 비율을 X 라고 할 때
2. 과속탐지 카메라에 적발되는 과속 차량들 사이의 시간간격을 X 라고 할 때

- 1번 풀이

- $0 \leq x \leq 1$

- 2번 풀이

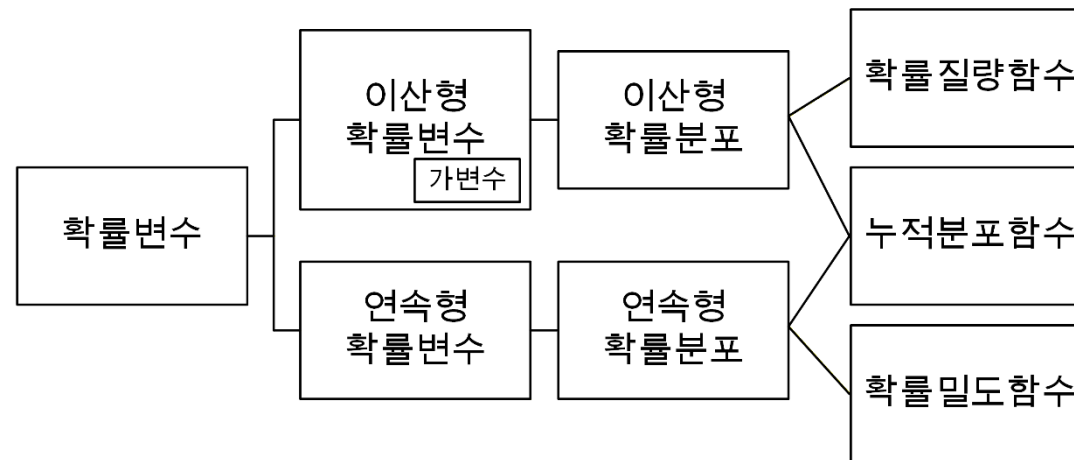
- $x \geq 0$

목 차

- 보충
- 확률변수의 개념
- 확률분포와 누적분포함수
- 결합확률분포

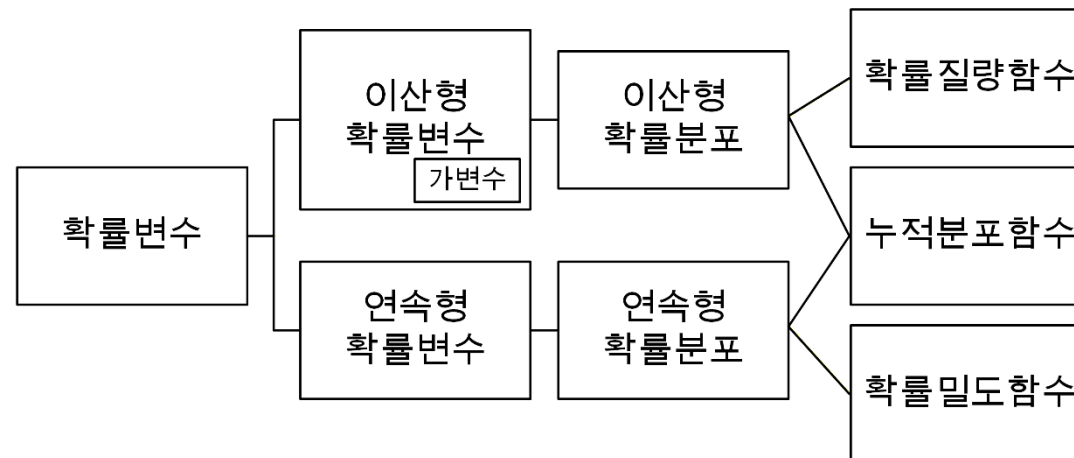
확률분포와 누적분포함수

- 확률분포(PD, Probability Distribution) (1/8)
 - 정의
 - 확률변수가 취할 수 있는 모든 값과 그 값들이 나타날 확률을 대응시켜 나타낸 관계
 - 구분
 - 이산형 확률분포(DPD, Discrete Probability Distribution)
 - 연속형 확률분포(CPD, Continuous Probability Distribution)
 - 결합확률분포(JPD, Joint Probability Distribution)



확률분포와 누적분포함수

- 확률분포(PD, Probability Distribution) (2/8)
 - 표현방식
 - 확률질량함수(PMF, Probability Mass Function)
 - 이산형 확률분포에서 사용
 - 확률밀도함수(PDF, Probability Density Function)
 - 연속형 확률분포에서 사용
 - 누적분포함수(CPF, Cumulative Probability Function)
 - 이산형, 연속형 확률분포에서 사용



확률분포와 누적분포함수

• 확률분포(PD, Probability Distribution) (3/8)

• 이산형 확률분포(DPD) (1/4)

• 정의

- 확률변수가 가질 수 있는 값이 명확하고 셀 수 있는 경우의 분포
 - e.g., IDS에 하루 동안 탐지되는 고위험군 악성 코드의 종류별 발생 분포

• 조건

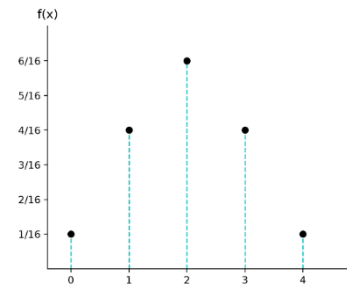
- $f(x) \geq 0$
- $\sum_x f(x) = 1$
- $P(X = x) = f(x)$

• 특징

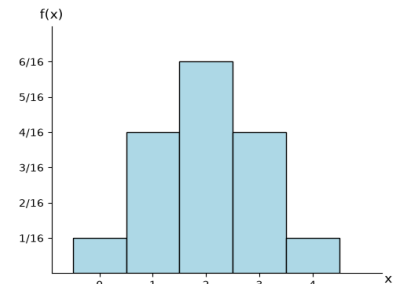
- 확률함수, 확률질량함수, 확률분포와 같은 명칭 사용
- 확률질량함수와 확률히스토그램을 활용한 시각적 표현 가능
 - 확률질량함수(PMF)
 - 확률 값을 x 축과 점선이나 실선으로 연결하여 이산형 확률 분포를 나타낸 그림
 - 확률히스토그램(Probability Histogram)
 - 확률질량함수도에서 점 대신 직사각형을 사용해 나타낸 그림

*IDS (Intrusion Detection System)

네트워크나 시스템의 비정상적인 활동을 감지해 관리자에게 알리는 침입 탐지 시스템.



<그림 1>
확률질량함수도



<그림 2>
확률히스토그램

확률분포와 누적분포함수

- 확률분포(PD, Probability Distribution) (4/8)

- 이산형 확률분포(DPD) (2/4)

- 예제 3.7

상점에 진열된 20대의 노트북 중에 불량품이 3대 끼여 있다. 어느 학교에서 이 중 임의로 2대를 구입하였을 때, 불량품의 개수의 확률분포를 구하라.

- 풀이

- 구입한 노트북 중 불량품의 수를 확률변수 X 라 하면,
 - X 의 값 x 가 취할 수 있는 값은 0, 1, 2

- $$f(0) = P(X = 0) = \frac{\binom{3}{0}\binom{17}{2}}{\binom{20}{2}} = \frac{\frac{3!}{0!3!} \times \frac{17!}{2!15!}}{\frac{20!}{2!18!}} = \frac{68}{95}$$
- $$f(1) = P(X = 1) = \frac{\binom{3}{1}\binom{17}{1}}{\binom{20}{2}} = \frac{\frac{3!}{1!2!} \times \frac{17!}{1!16!}}{\frac{20!}{2!18!}} = \frac{51}{190}$$
- $$f(2) = P(X = 2) = \frac{\binom{3}{2}\binom{17}{0}}{\binom{20}{2}} = \frac{\frac{3!}{2!1!} \times \frac{17!}{0!17!}}{\frac{20!}{2!18!}} = \frac{3}{190}$$

x	0	1	2
$f(x)$	$\frac{68}{95}$	$\frac{51}{190}$	$\frac{3}{190}$

확률분포와 누적분포함수

- 확률분포(PD, Probability Distribution) (5/8)

- 이산형 확률분포(DPD) (3/4)

- 예제 3.8 (1/2)

어느 대리점에서 판매된 외제차의 50%에 디젤엔진이 장착되었다고 할 때, 이 대리점에서 다음달에 판매될 4대의 외제차 가운데 디젤엔진이 장착된 차의 수의 확률분포에 대한 식을 구하라.

- 풀이

- 디젤모델이나 가솔린모델을 판매할 확률은 동일하게 $\frac{1}{2}$
 - 표본공간에는 발생확률이 동일한 $2^4 = 16$ 개의 표본점 존재

- 4대 중 디젤엔진인 차량의 수를 확률변수 X
 - X 가 가질 수 있는 값 x 는 0, 1, 2, 3, 4

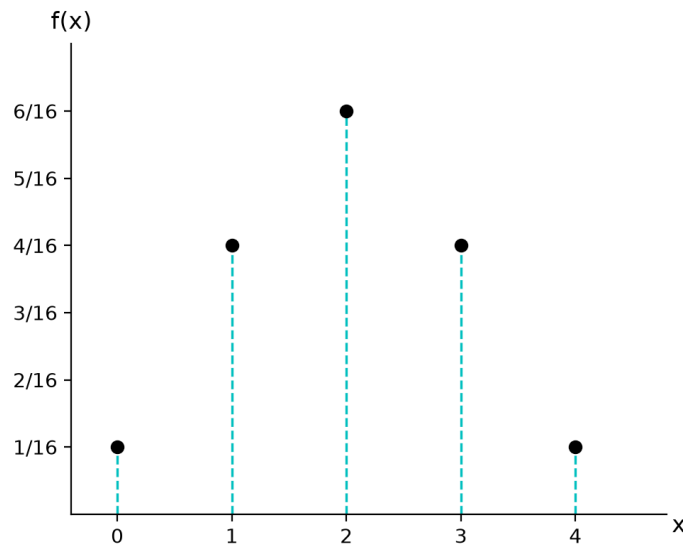
- 디젤엔진일 확률을 $\binom{4}{x}$ 로 표현 가능

∴ 확률분포 $f(x) = \frac{\binom{4}{x}}{16}, x = 0, 1, 2, 3, 4$

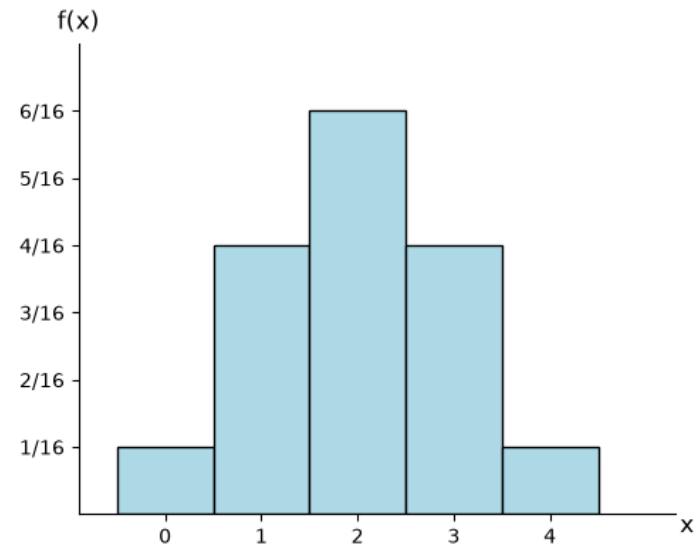
확률분포와 누적분포함수

- 확률분포(PD, Probability Distribution) (6/8)
 - 이산형 확률분포(DPD) (4/4)
 - 예제 3.8 (2/2)

어느 대리점에서 판매된 외제차의 50%에 디젤엔진이 장착되었다고 할 때, 이 대리점에서 다음달에 판매될 4대의 외제차 가운데 디젤엔진이 장착된 차의 수의 확률분포에 대한 식을 구하라.



<그림 1>
확률질량함수도



<그림 2>
확률히스토그램

확률분포와 누적분포함수

- 확률분포(PD, Probability Distribution) (7/8)

- 연속형 확률분포(CPD) (1/2)

- 정의

- 확률변수가 소수점 아래로 무한히 쪼개지는 연속적인 값을 가질 때의 확률분포

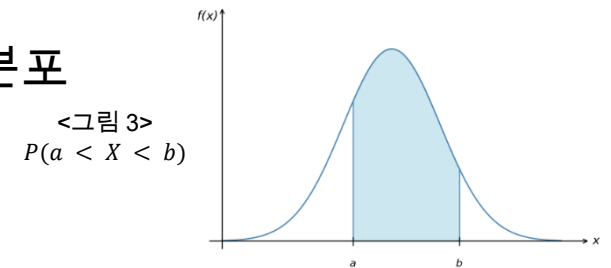
- e.g., 암호화 연산 중 칩의 실시간 전력 소모량 분포

- 조건

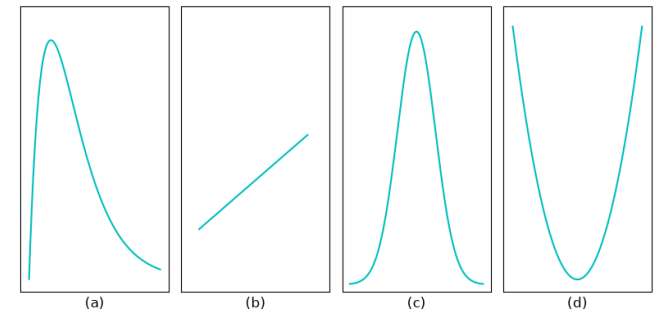
- 모든 $x \in R$ 에 대해 $f(x) \geq 0$
 - $\int_{-\infty}^{\infty} f(x) dx = 1$
 - $P(a < X < b) = \int_a^b f(x) dx$

- 특징

- 이산형 확률분포처럼 표형태 불가능
 - 전체 면적이 1인 확률밀도함수 사용
 - 유한개의 불연속점을 갖는 것이 가능



<그림 4>
전형적인 밀도함수



확률분포와 누적분포함수

- 확률분포(PD, Probability Distribution) (8/8)

- 연속형 확률분포(CPD) (2/2)

- 예제 3.9

제어실험에서 반응온도(°C)의 변화에 따른 오차는 다음과 같은 확률분포를 가지는 연속확률 변수 X 라고 가정하자.

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{그 외의 경우} \end{cases}$$

(a) 함수 $f(x)$ 가 확률밀도함수임을 증명하라.

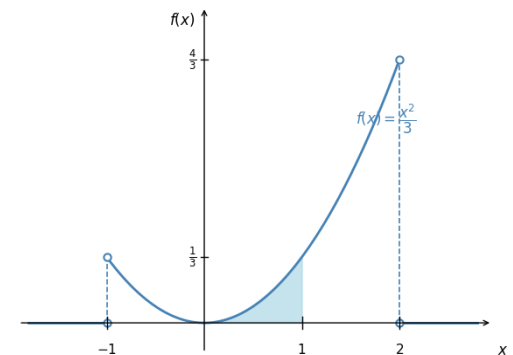
(b) 확률 $P(0 < X \leq 1)$ 을 구하라.

- 풀이

- (a) $f(x) \geq 0$ 임은 명확하므로, 다음 식을 통해 확률밀도함수임을 보일 수 있음

- $\int_{-\infty}^{\infty} f(x) dx = \int_{-1}^2 \frac{x^2}{3} dx = \frac{x^3}{9} \Big|_{-1}^2 = \frac{8}{9} + \frac{1}{9} = 1$

- (b) $P(0 < x \leq 1) = \int_0^1 \frac{x^2}{3} dx = \frac{x^3}{9} \Big|_0^1 = \frac{1}{9}$



<그림 5>
확률밀도함수

확률분포와 누적분포함수

- 누적분포함수(CDF, Cumulative Distribution Function) (1/6)
 - 정의
 - 실수 x 에 대해 확률변수 X 가 x 이하일 확률을 나타내는 함수
 - e.g., 로그인 실패 횟수가 5회 이하일 누적 확률
 - 특징
 - 비감소 함수, 주어진 확률변수 뿐만 아니라 모든 실수에서 정의
 - 이산형 확률분포에서의 누적분포함수 $F(x)$
 - $F(x) = P(X \leq x) = \sum_{t \leq x} f(t), -\infty < x < \infty$
 - 연속형 확률분포에서의 누적분포함수 $F(x)$
 - $F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt, -\infty < x < \infty$
 - 필요성
 - 연산의 단순화
 - 프로그래밍/알고리즘 표준화

확률분포와 누적분포함수

- 누적분포함수(CDF, Cumulative Distribution Function) (2/6)

- 예제 3.10 (1/3)

예제 3.8에서 확률변수 X 의 확률분포는 다음과 같다. 이때, 확률변수 X 의 누적분포 $F(x)$ 를 구하라. 그리고 $F(x)$ 를 사용하여 $f(2) = \frac{3}{8}$ 이 됨을 증명하라.

x	0	1	2	3	4
$f(x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

- 풀이

- 다음 문제에서 주어진 표를 바탕으로 누적분포를 구해볼 때,

- $F(0) = f(0) = \frac{1}{16}$

- $F(1) = f(0) + f(1) = \frac{5}{16}$

- $F(2) = f(0) + f(1) + f(2) = \frac{11}{16}$

- $F(3) = f(0) + f(1) + f(2) + f(3) = \frac{15}{16}$

- $F(4) = f(0) + f(1) + f(2) + f(3) + f(4) = 1$

$$\therefore F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{16}, & 0 \leq x < 1 \\ \frac{5}{16}, & 1 \leq x < 2 \\ \frac{11}{16}, & 2 \leq x < 3 \\ \frac{15}{16}, & 3 \leq x < 4 \\ 1, & x \geq 4 \end{cases}$$
$$\therefore f(2) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{3}{8}$$

확률분포와 누적분포함수

- 누적분포함수(CDF, Cumulative Distribution Function) (3/6)

- 예제 3.10 (2/3)

예제 3.8에서 확률변수 X 의 확률분포는 다음과 같다. 이때, 확률변수 X 의 누적분포 $F(x)$ 를 구하라. 그리고 $F(x)$ 를 사용하여 $f(2) = \frac{3}{8}$ 이 됨을 증명하라.

x	0	1	2	3	4
$f(x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

- 풀이

$$\therefore F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{16}, & 0 \leq x < 1 \\ \frac{5}{16}, & 1 \leq x < 2 \\ \frac{11}{16}, & 2 \leq x < 3 \\ \frac{15}{16}, & 3 \leq x < 4 \\ 1, & x \geq 4 \end{cases}$$

$$\therefore f(2) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{3}{8}$$

확률분포와 누적분포함수

- 누적분포함수(CDF, Cumulative Distribution Function) (4/6)

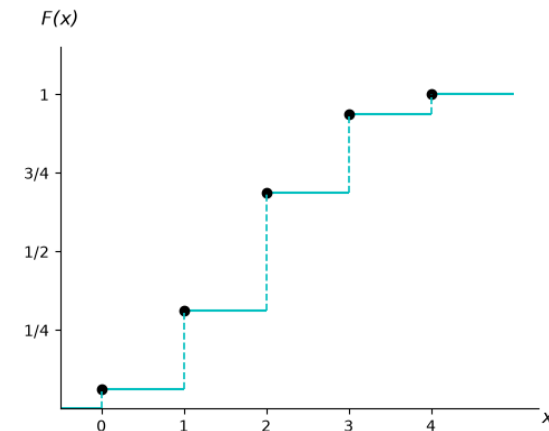
- 예제 3.10 (3/3)

예제 3.8에서 확률변수 X 의 확률분포는 다음과 같다. 이때, 확률변수 X 의 누적분포 $F(x)$ 를 구하라. 그리고 $F(x)$ 를 사용하여 $f(2) = \frac{3}{8}$ 이 됨을 증명하라.

x	0	1	2	3	4
$f(x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

- 풀이

$$\therefore F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{16}, & 0 \leq x < 1 \\ \frac{5}{16}, & 1 \leq x < 2 \\ \frac{11}{16}, & 2 \leq x < 3 \\ \frac{15}{16}, & 3 \leq x < 4 \\ 1, & x \geq 4 \end{cases}$$



<그림 6>
이산형 누적분포

확률분포와 누적분포함수

- 누적분포함수(CDF, Cumulative Distribution Function) (5/6)

- 예제 3.11

예제 3.9의 확률밀도함수 $f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{그 외의 경우} \end{cases}$ 에 대해 누적분포 $F(x)$ 를 구하고, 그것을 이용하여 $P(0 < X \leq 1)$ 을 구하라.

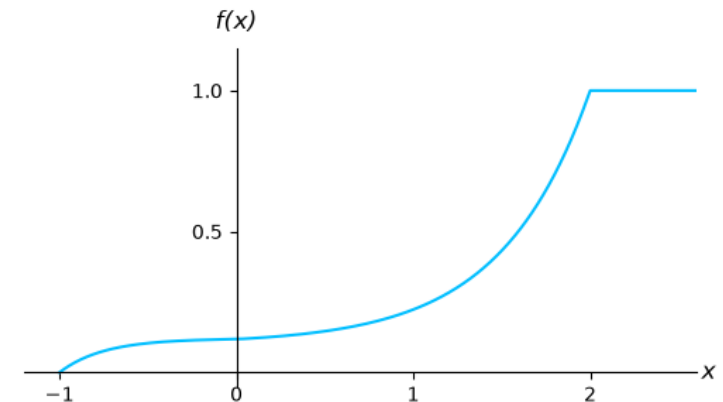
- 풀이

- $-1 < x < 2$ 에 대해

- $F(x) = \int_{-\infty}^x f(t) dt = \int_{-1}^x \frac{t^2}{3} dt = \frac{t^3}{9} \Big|_{-1}^x = \frac{x^3+1}{9}$

- $\therefore F(x) = \begin{cases} 0, & x < -1 \\ \frac{x^3+1}{9}, & -1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$

$$\therefore P(0 < X \leq 1) = F(1) - F(0) = \frac{2}{9} - \frac{1}{9} = \frac{1}{9}$$



<그림 7>
연속형 누적분포

확률분포와 누적분포함수

- 누적분포함수(CDF, Cumulative Distribution Function) (6/6)

- 예제 3.12

어느 기관에서 프로젝트를 입찰에 부치고 적절한 입찰가를 b 로 예상한다고 할 때, 해당 기관은 낙찰가 y 의 밀도함수를 다음과 같이 생각하고 있다.

$$f(y) = \begin{cases} \frac{5}{8b}, & \frac{2}{5}b \leq y \leq 2b \\ 0, & \text{그 외의 경우} \end{cases}$$

$F(y)$ 를 구하고, 이것을 이용하여 낙찰가 y 가 해당 기관의 예상치 b 보다 작을 확률을 구하라.

- 풀이

- $\frac{2b}{5} \leq y \leq 2b$

- $F(y) = \int_{\frac{2b}{5}}^y \frac{5}{8b} dt = \frac{5t}{8b} \Big|_{\frac{2b}{5}}^y = \frac{5y}{8b} - \frac{1}{4}$

- $F(y) = \begin{cases} 0, & y < \frac{2}{5}b \\ \frac{5y}{8b} - \frac{1}{4}, & \frac{2}{5}b \leq y < 2b \\ 1, & y \geq 2b \end{cases}$

- 낙찰가가 예상치 b 보다 작을 확률은 $P(Y \leq b) = F(b) = \frac{5}{8} - \frac{1}{4} = \frac{3}{8}$

확률분포와 누적분포함수

- 확률분포 시각화 (1/5)

- 확률질량함수도

- 정의

- 확률 값을 x 축과 점선이나 실선으로 연결하여 이산형 확률 분포를 나타낸 그림

- 장점

- 개별 사건의 정확한 확률을 직관적으로 확인 가능

- 단점

- 연속적인 값에는 적용 불가

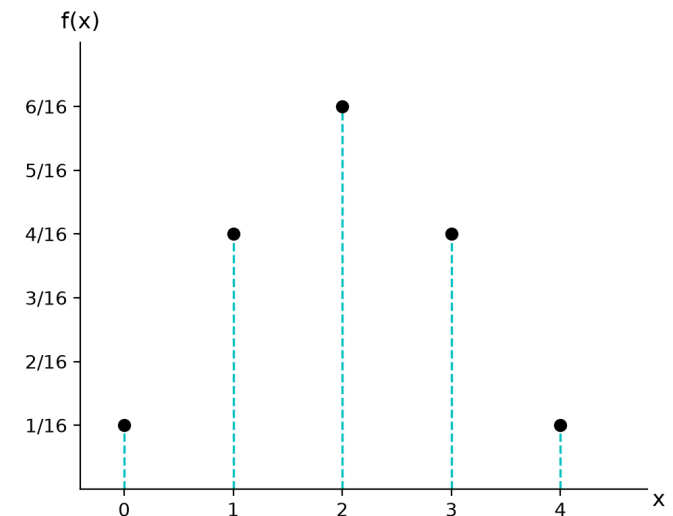
- 활용 예시

- 무차별 대입 공격 탐지

- e.g., 특정 IP가 1분 동안 로그인 실패를 5회, 10회, 15회 할 확률 계산

- 디지털 포렌식

- e.g., 악성코드 파일 내에서 특정 바이트가 나타나는 빈도 분석



<그림 1>
확률질량함수도

확률분포와 누적분포함수

- 확률분포 그래프 (2/5)

- 확률히스토그램

- 정의

- 확률질량함수도에서 점 대신 직사각형을 사용해 나타낸 그림

- 장점

- 데이터의 밀집도 파악 가능

- 단점

- 구간 설정에 따라 그래프의 모양이 왜곡될 수 있음

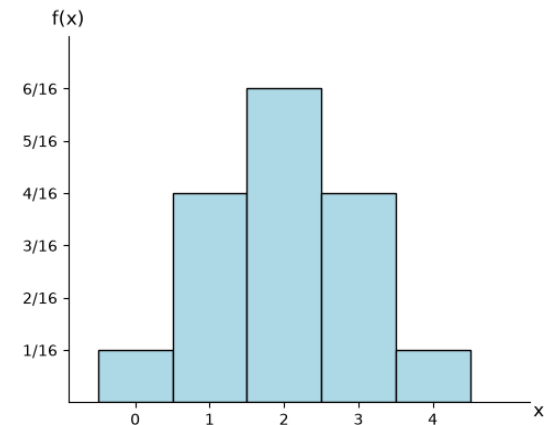
- 활용 예시

- 트래픽 베이스라인 모니터링

- e.g., 사내 망의 시간대별 패킷 크기 분포를 시각화하여, 평소와 다른 기형적인 모양이 나타나는지 감시

- 보안 장비 리소스 모니터링

- e.g., 방화벽 CPU 사용량 분포 시각화



<그림 2>
확률히스토그램

확률분포와 누적분포함수

- 확률분포 그래프 (3/5)

- 확률밀도함수

- 정의

- 특정 지점의 y 축 값은 확률이 아닌 밀도이며, 특정 구간의 면적이 확률인 그림

- 장점

- 데이터가 밀집된 구간과 희소한 구간을 매끄러운 곡선으로 표현 가능

- 단점

- 확률 계산 시, 적분을 해야 하므로 번거로움 발생

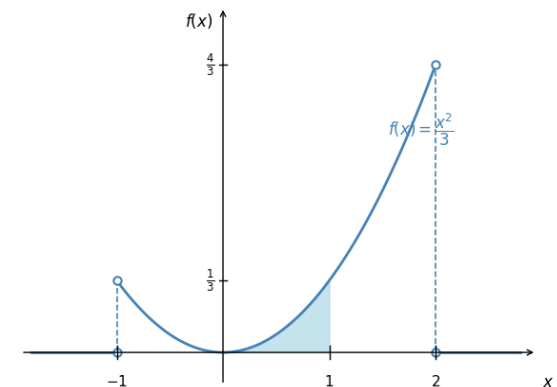
- 활용 예시

- 사용자 및 개체 행동 분석

- e.g., 특정 임직원의 평소 시스템 접속 시간 분포 곡선 생성 후, 해당 이상 징후 탐지

- APT 공격 탐지

- e.g., C&C 서버와 통신하는 주기의 밀도 분석



<그림 5>
확률밀도함수

확률분포와 누적분포함수

- 확률분포 그래프 (4/5)

- 누적분포함수 (1/2)

- 정의

- 실수 x 에 대해 확률변수 X 가 x 이하일 확률을 나타내는 함수

- 장점

- 적분 없이 단순 뺄셈을 통해 구간의 확률 계산 가능

- 단점

- 확률이 누적되기 때문에, 데이터가 어느 특정

- 활용 예시

- 리스크 관리 및 가설 검정

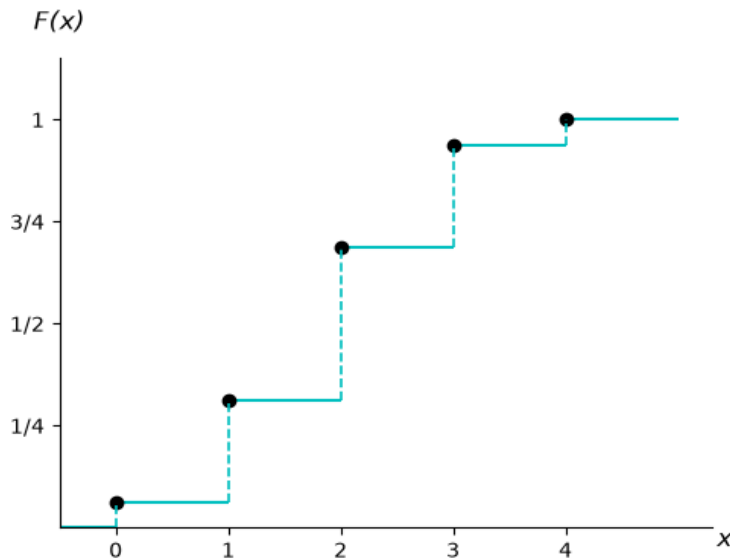
- e.g., 위험 평가 시 “자산이 랜섬웨어에 감염되어 발생할 손실액이 1억 원 이하일 경우”의 누적 확률 계산

- 실시간 탐지 알람 제어

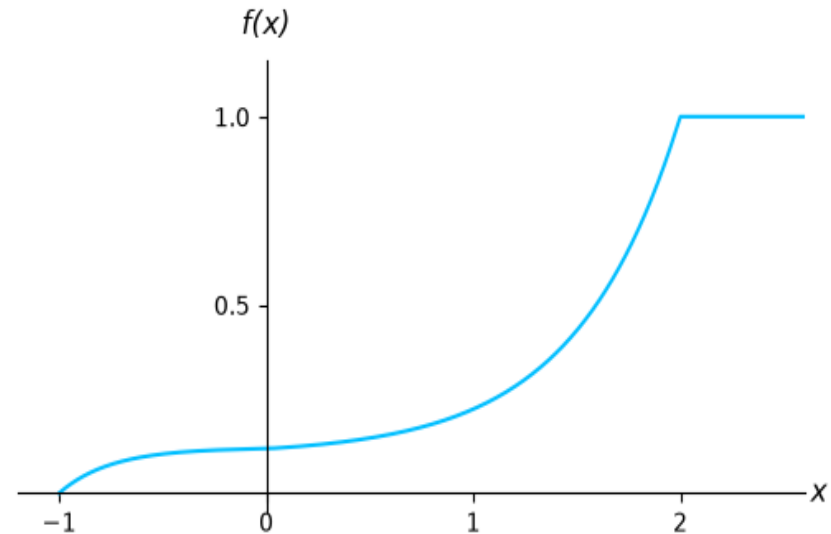
- e.g., 보안 솔루션이 탐지한 위협 점수를 누적분포함수에 대입하여 “현재 이벤트가 상위 0.1%에 해당하는 극단적 위협”인지 즉시 판정하여 차단 알림 전송

확률분포와 누적분포함수

- 확률분포 그래프 (5/5)
- 누적분포함수 (2/2)



<그림 6>
이산형 누적분포



<그림 7>
연속형 누적분포

목 차

- 보충
- 확률변수의 개념
- 확률분포와 누적분포함수
- 결합확률분포

결합 확률 분포

• 결합 확률 분포 (Joint Probability Distribution) (1/6)

• 정의

- 두 개 이상의 확률변수가 동시에 어떤 값을 가질 확률을 나타내는 분포
 - e.g., 로그인 시도 횟수와 악성 파일 다운로드 수가 동시에 특정 값일 확률

• 필요성

- 여러 개의 확률변수들의 결과를 동시에 취해야 할 때 사용
 - e.g., 어떤 서버에 대한 공격 탐지에서, 초당 패킷 수 X 와 비정상 포트 접근 시도 횟수 Y 의 결과는 (x, y) 로 동시에 표시

• 표현

- 결합 확률 질량 함수 (JPMF, Joint Probability Mass Function)
 - 확률변수 X, Y 가 이산형 확률변수일 경우
- 결합 밀도 함수 (JDF, Joint Density Function)
 - 확률변수 X, Y 가 연속형 확률변수일 경우

결합 확률 분포

- 결합 확률 분포 (Joint Probability Distribution) (2/6)
- 결합 확률 질량 함수 (JPMF) (1/2)
 - 정의
 - 두 개 이상의 이산형 확률변수 X, Y 가 각각 특정 값 x, y 를 동시에 가질 확률을 대응시키는 함수
 - e.g., 로그인 실패 횟수와 악성 파일 다운로드 개수가 각각 5회와 2개일 확률
 - 조건
 - 모든 (x, y) 에 대해 $f(x, y) \geq 0$
 - $\sum_x \sum_y f(x, y) = 1$
 - $P(X = x, Y = y) = f(x, y)$
 - x, y 평면상의 어떤 영역 A 에 대하여 $P[(X, Y) \in A] = \sum \sum_A f(x, y)$

결합확률분포

• 결합확률분포(Joint Probability Distribution) (3/6)

• 결합확률질량함수(JPMF) (2/2)

• 예제 3.13

3개의 청색, 2개의 적색, 3개의 녹색 볼펜이 들어있는 상자에서 임의로 2개를 추출하고자 한다. X 를 청색 볼펜의 수, Y 를 적색 볼펜의 수라고 할 때,

(a) 결합확률분포 $f(x, y)$ 를 구하라.

(b) $A = \{(x, y) | x + y \leq 1\}$ 라고 할 때, $P[(X, Y) \in A]$ 를 구하라.

• 풀이

• (a) 결합확률분포 $f(x, y)$

• 8개 볼펜 중 임의의 2개를 추출하는 경우의 수: $\binom{8}{2}$

• 청색 볼펜을 뽑는 경우의 수: $\binom{3}{x}$

• 적색 볼펜을 뽑는 경우의 수: $\binom{2}{y}$

• 녹색 볼펜을 뽑는 경우의 수: $\binom{3}{2-x-y}$

$$\therefore f(x, y) = \frac{\binom{3}{x} \binom{2}{y} \binom{3}{2-x-y}}{\binom{8}{2}}, \quad x = 0, 1, 2, y = 0, 1, 2, 0 \leq x + y \leq 2$$

• (b) $P[(X, Y) \in A] = P(X + Y \leq 1) = f(0, 0) + f(0, 1) + f(1, 0) = \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}$

$f(x, y)$		x			행의 합
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$		$\frac{3}{7}$
	2	$\frac{1}{28}$			$\frac{1}{28}$
열의 합		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

결합 확률 분포

- 결합 확률 분포(Joint Probability Distribution) (4/6)
- 결합 밀도 함수(JDF) (1/2)
 - 정의
 - 두 개 이상의 연속형 확률변수 X, Y 가 연속적인 공간에서 동시에 특정 구간에 존재할 확률의 밀도를 나타내는 함수
 - e.g., 웹 서버의 응답 지연 시간이 1~2초 사이이면서 CPU 실시간 사용률이 80~90% 사이일 확률 밀도
 - 조건
 - 모든 (x, y) 에 대해 $f(x, y) \geq 0$
 - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1$
 - $P[(X, Y) \in A] = \int \int_A f(x, y) \, dx \, dy$ (A 는 xy 평면 상의 임의의 영역)

결합 확률 분포

- 결합 확률 분포 (Joint Probability Distribution) (5/6)

- 결합 밀도 함수 (JDF) (2/2)

- 예제 3.14 (1/2)

어느 과자회사에서는 연한 초콜릿과 진한 초콜릿을 입힌 과자상자를 취급하고 있다. 임의로 하나의 과자상자를 선택했을 때 X 와 Y 를 각각 연한 초콜릿과 진한 초콜릿의 비율이라 하면, 결합 확률 분포는 다음과 같이 주어진다.

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{그 외의 경우} \end{cases}$$

이때 다음을 구하라.

(a) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ 을 증명하라.

- (a) 풀이

- $$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_0^1 \int_0^1 \frac{2}{5}(2x + 3y) dx dy = \int_0^1 \left(\frac{2x^2}{5} + \frac{6xy}{5} \right) \Big|_{x=0}^{x=1} dy \\ &= \int_0^1 \left(\frac{2}{5} + \frac{6y}{5} \right) dy = \left(\frac{2y}{5} + \frac{3y^2}{5} \right) \Big|_0^1 = \frac{2}{5} + \frac{3}{5} = 1 \end{aligned}$$

결합확률분포

- 결합확률분포(Joint Probability Distribution) (6/6)

- 결합밀도함수(JDF) (2/2)

- 예제 3.14 (2/2)

어느 과자회사에서는 연한 초콜릿과 진한 초콜릿을 입힌 과자상자를 취급하고 있다. 임의로 하나의 과자상자를 선택했을 때 X 와 Y 를 각각 연한 초콜릿과 진한 초콜릿의 비율이라 하면, 결합확률분포는 다음과 같이 주어진다.

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{그 외의 경우} \end{cases}$$

이때 다음을 구하라.

(b) $A = \{(x, y) | 0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$ 일 때의 $P[(X, Y) \in A]$ 를 구하라.

- (b) 풀이

$$\begin{aligned} P[(X, Y) \in A] &= P\left(0 < X < \frac{1}{2}, \frac{1}{4} < Y < \frac{1}{2}\right) = \int_{\frac{1}{4}}^{\frac{1}{2}} \int_0^{\frac{1}{2}} \frac{2}{5}(2x + 3y) \, dx \, dy = \int_{\frac{1}{4}}^{\frac{1}{2}} \left(\frac{2x^2}{5} + \frac{6xy}{5}\right) \Big|_{x=0}^{x=\frac{1}{2}} dy \\ &= \int_{\frac{1}{4}}^{\frac{1}{2}} \left(\frac{1}{10} + \frac{3y}{5}\right) dy = \left(\frac{y}{10} + \frac{3y^2}{10}\right) \Big|_{\frac{1}{4}}^{\frac{1}{2}} = \frac{1}{10} \left[\left(\frac{1}{2} + \frac{3}{4}\right) - \left(\frac{1}{4} + \frac{3}{16}\right)\right] = \frac{1}{10} \left(\frac{5}{4} - \frac{7}{16}\right) = \frac{13}{160} \end{aligned}$$

결합 확률 분포

- 주변분포(Marginal Distribution) (1/4)

- 정의

- 결합분포에서 특정 변수만을 대상으로 나머지 변수의 영향을 제거한 분포

- 구분

- 이산형 확률변수 X, Y

- $g(x) = \sum_y f(x, y), \quad h(y) = \sum_x f(x, y)$

- 연속형 확률변수 X, Y

- $g(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad h(y) = \int_{-\infty}^{\infty} f(x, y) dx$

- 특징 (1/2)

- 확률변수가 n 개 일 경우, 특정 확률 변수의 주변분포를 다음과 같이 계산 가능

- 이산형인 경우, $g(x_1) = \sum_{x_2} \cdots \sum_{x_n} f(x_1, x_2, \cdots, x_n)$

- 연속형인 경우, $g(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n) dx_2 dx_3 \cdots dx_n$

결합 확률 분포

- 주변분포(Marginal Distribution) (2/4)

- 특징 (2/2)

- 확률 변수가 n 개 일 경우, 특정 변수가 2개에 대한 결합주변 분포(Joint Conditional Distribution)로 확장하여 사용 가능
 - 이산형인 경우, $g(x_1, x_2) = \sum_{x_3} \cdots \sum_{x_n} f(x_1, x_2, \cdots, x_n)$
 - 연속형인 경우, $g(x_1, x_2) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n) dx_3 dx_4 \cdots dx_n$

- 필요성

- 여러 확률변수 중 특정 변수의 정보만 추출할 때 사용
 - e.g., 로그인 시도 횟수 X , 사용자 권한등급 Y , 접속 IP 대역 Z 의 결합 분포에서 권한등급 Y 의 분포만 얻고자 하는 경우
- 분석 이전에 각 변수의 개별 분포 파악이 필요할 때 사용
 - e.g., 패킷 크기 X 에 따른 탐지 여부 Y 분석 이전에, X 와 Y 각각의 분포를 먼저 확인하고자 하는 경우

결합 확률 분포

- 주변분포(Marginal Distribution) (3/4)

- 예제 3.15

아래 표의 열과 행의 합이 각각 X, Y 의 주변분포가 됨을 증명하라.

$f(x, y)$		x			행의 합
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$		$\frac{3}{7}$
	2	$\frac{1}{28}$			$\frac{1}{28}$
열의 합		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

- 풀이

- 확률변수 X 에 대해 $g(x)$ 를 X 의 주변분포, $h(y)$ 를 Y 의 주변분포라고 할 때,

- $g(0) = f(0,0) + f(0,1) + f(0,2) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14}$

- $g(1) = f(1,0) + f(1,1) + f(1,2) = \frac{9}{28} + \frac{3}{14} + 0 = \frac{15}{28}$

- $g(2) = f(2,0) + f(2,1) + f(2,2) = \frac{3}{28} + 0 + 0 = \frac{3}{28}$

위 결과를 통해 문제에서 주어진 표의 열의 합과 동일함을 알 수 있으며, 동일한 방법으로 $h(y)$ 의 값들이 행의 합과 일치함을 알 수 있음

결합 확률 분포

- 주변분포(Marginal Distribution) (4/4)

- 예제 3.16

예제 3.14의 결합분포 $f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{그 외의 경우} \end{cases}$ 에 대해 $g(x)$ 와 $h(y)$ 를 구하라.

- 풀이

- $0 \leq x \leq 1$ 구간에서

- $g(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 \frac{2}{5}(2x + 3y) dy = \left(\frac{4xy}{5} + \frac{6y^2}{10} \right) \Big|_{y=0}^{y=1} = \frac{4x+3}{5}$
 - 그 외의 경우에는 $g(x) = 0$ 이 됨

- $0 \leq y \leq 1$ 구간에서

- $h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 \frac{2}{5}(2x + 3y) dx = \left(\frac{2x^2}{5} + \frac{6xy}{5} \right) \Big|_{x=0}^{x=1} = \frac{2(1+3y)}{5}$
 - 그 외의 경우에는 $h(y) = 0$ 이 됨

결합 확률 분포

- 조건부분포(Conditional Distribution) (1/6)

- 정의

- 결합분포에서 한 확률변수의 값을 고정했을 때, 나머지 확률변수가 따르는 분포

- 구분

- $X = x$ 로 주어졌을 때 확률변수 Y 의 조건부분포

- $f(y|x) = \frac{f(x,y)}{g(x)}, (g(x) > 0)$

- $Y = y$ 로 주어졌을 때 확률변수 X 의 조건부분포

- $f(x|y) = \frac{f(x,y)}{h(y)}, (h(y) > 0)$

- 특징 (1/2)

- 두 개 이상의 확률변수에 대한 결합조건부분포(Joint Conditional Distribution)로 확장하여 사용 가능

- $f(x_1, x_2, \dots, x_{a-1} | x_a, x_{a+1}, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n)}{g(x_a, x_{a+1}, \dots, x_n)}$

결합 확률 분포

• 조건부분포(Conditional Distribution) (2/6)

• 특징 (2/2)

- $Y = y$ 일 때 X 가 $a < X < b$ 일 확률은 다음과 같이 계산 가능
 - 이산형인 경우, $P(a < X < b | Y = y) = \sum_{a < X < b} f(x|y)$
 - 연속형인 경우, $P(a < X < b | Y = y) = \int_a^b f(x|y) dx$

• 필요성

- 특정 조건이 주어졌을 때 나머지 변수의 분포의 변화를 파악할 때 사용
 - e.g., 접속 IP 대역 X 가 해외로 확인된 경우, 로그인 시도 횟수 Y 의 분포가 어떻게 변하는지 분석
- 두 변수 간 의존관계 여부를 판단할 때 사용
 - e.g., 사용자 권한등급 X 가 주어졌을 때 비정상 행위 탐지 여부 Y 의 분포가 달라지는지를 통해, 두 변수 간 연관성 존재 여부를 확인

결합 확률 분포

• 조건부분포(Conditional Distribution) (3/6)

• 예제 3.17

예제 3.13에서 $Y = 1$ 로 주어졌을 때 X 의 조건부분포를 구하고, 그것을 이용하여 $P(X = 0|Y = 1)$ 을 구하라.

• 풀이

- $y = 1$ 일 때의 $f(x|y)$ 가 필요하므로 $h(1)$ 을 구해야 함
- $h(1) = \sum_{x=0}^2 f(x, 1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}$ 이므로
- $f(x|1) = \frac{f(x,1)}{h(1)} = \frac{7}{3} f(x, 1)$, $x = 0, 1, 2$ 이 됨
- $f(0|1) = \left(\frac{7}{3}\right) f(0,1) = \left(\frac{7}{3}\right) \left(\frac{3}{14}\right) = \frac{1}{2}$,
- $f(1|1) = \left(\frac{7}{3}\right) f(1,1) = \left(\frac{7}{3}\right) \left(\frac{3}{14}\right) = \frac{1}{2}$,
- $f(2|1) = \left(\frac{7}{3}\right) f(2,1) = \left(\frac{7}{3}\right) (0) = 0$
- $P(X = 0|Y = 1) = f(0|1) = \frac{1}{2}$

$f(x, y)$		x			행의 합
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$		$\frac{3}{7}$
	2	$\frac{1}{28}$			$\frac{1}{28}$
열의 합		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

- 두 개의 볼펜 중 하나가 적색이라는 사실이 알려지면 다른 하나의 볼펜이 청색이 아닐 확률은 $\frac{1}{2}$ 가 됨을 알 수 있음

결합 확률 분포

• 조건부 분포(Conditional Distribution) (4/6)

• 예제 3.18 (1/2)

X 와 Y 를 각각 단위 온도 변화량과 어떤 원자가 방출하는 스펙트럼 변화율을 나타내는 확률변수라 할 때, 확률변수 (X, Y) 에 대한 결합밀도함수는 다음과 같다.

$$f(x, y) = \begin{cases} 10xy^2, & 0 < x < y < 1 \\ 0, & \text{그 외의 경우} \end{cases}$$

(a) 주변밀도함수 $g(x), h(y)$ 와 조건부밀도함수 $f(y|x)$ 를 구하라.

• (a) 풀이

$$\bullet \quad g(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_x^1 10xy^2 dy = \frac{10}{3}xy^3 \Big|_{y=x}^{y=1} = \frac{10}{3}x(1 - x^3), \quad 0 < x < 1$$

$$\bullet \quad h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^y 10xy^2 dx = 5x^2y^2 \Big|_{x=0}^{x=y} = 5y^4, \quad 0 < y < 1$$

$$\bullet \quad f(y|x) = \frac{f(x, y)}{g(x)} = \frac{10xy^2}{\frac{10}{3}x(1-x^3)} = \frac{3y^2}{1-x^3}, \quad 0 < x < y < 1$$

결합 확률 분포

- 조건부 분포(Conditional Distribution) (5/6)

- 예제 3.18 (2/2)

X 와 Y 를 각각 단위 온도 변화량과 어떤 원자가 방출하는 스펙트럼 변화율을 나타내는 확률변수라 할 때, 확률변수 (X, Y) 에 대한 결합 밀도 함수는 다음과 같다.

$$f(x, y) = \begin{cases} 10xy^2, & 0 < x < y < 1 \\ 0, & \text{그 외의 경우} \end{cases}$$

(b) 온도가 0.25 단위 높아졌을 때 스펙트럼 변화량이 $\frac{1}{2}$ 보다 클 확률을 구하라.

- (b) 풀이

- $P\left(Y > \frac{1}{2} \mid X = 0.25\right) = \int_{\frac{1}{2}}^1 f(y \mid x = 0.25) dy = \int_{\frac{1}{2}}^1 \frac{3y^2}{1-0.25^3} dy = \frac{64}{63} y^3 \Big|_{\frac{1}{2}}^1 = \frac{8}{9}$

결합확률분포

• 조건부분포(Conditional Distribution) (6/6)

• 예제 3.19

결합확률분포가 다음과 같이 주어졌을 때 $g(x), h(y), f(x|y)$ 를 구하고, $P(\frac{1}{4} < X < \frac{1}{2} | Y = \frac{1}{3})$ 의 값을 구하라.

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2, 0 < y < 1 \\ 0, & \text{그 외의 경우} \end{cases}$$

• 풀이

- $g(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 \frac{x(1+3y^2)}{4} dy = \left(\frac{xy}{4} + \frac{xy^3}{4} \right) \Big|_{y=0}^{y=1} = \frac{x}{2}, 0 < x < 2$
- $h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^2 \frac{x(1+3y^2)}{4} dx = \left(\frac{x^2}{8} + \frac{3x^2y^2}{8} \right) \Big|_{x=0}^{x=2} = \frac{1+3y^2}{2}, 0 < y < 1$
- $f(x|y) = \frac{f(x, y)}{h(y)} = \frac{\frac{x(1+3y^2)}{4}}{\frac{1+3y^2}{2}} = \frac{x}{2}, 0 < x < 2$
- $P\left(\frac{1}{4} < X < \frac{1}{2} \mid Y = \frac{1}{3}\right) = \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{x}{2} dx = \frac{x^2}{4} \Big|_{\frac{1}{4}}^{\frac{1}{2}} = \frac{3}{64}$

결합 확률 분포

• 통계적 독립(Statistically Independent) (1/3)

• 정의

*독립(Independent)

어떠한 사건이 다른 사건에 영향을 미치지 않는 것.

- X 와 Y 를 결합 확률 분포 $f(x, y)$ 와 주변 분포 $g(x), h(y)$ 를 가지는 이산형 혹은 연속형 확률 변수라 할 때, 아래의 조건을 만족하는 확률 변수 X 와 Y 의 관계

- 모든 (x, y) 에 대하여, $f(x, y) = g(x)h(y)$

• 특징

- 각각의 확률 변수가 서로에게 영향을 주거나 받지 않음
- 두 개 이상의 확률 변수에 대한 통계적 독립은 상호 통계적 독립(Matual Independent)이라 부름
 - $X_1, X_2, \dots, X_n$ 을 결합 확률 분포 $f(x_1, x_2, \dots, x_n)$ 과 주변 분포 $f_1(x_1), f_2(x_2), \dots, f_n(x_n)$ 을 가지는 이산형 혹은 연속형 확률 변수라 할 때, 모든 $(x_1, x_2, \dots, x_n)$ 에 대해 $f(x_1, x_2, \dots, x_n) = f_1(x_1)f_2(x_2)\dots f_n(x_n)$ 인 관계

결합 확률 분포

• 통계적 독립(Statistically Independent) (2/3)

• 예제 3.20

예제 3.13의 확률변수들이 통계적으로 독립이 아님을 증명하라.

• 풀이

• 점 $(0, 1)$ 을 고려해 볼 때, 오른쪽 표로부터 $f(0, 1), g(0), h(1)$ 을 구할 수 있음

$$• f(0, 1) = \frac{3}{14}$$

$$• g(0) = \sum_{y=0}^2 f(0, y) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14}$$

$$• h(1) = \sum_{x=0}^2 f(x, 1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}$$

$$• f(0, 1) \neq g(0)h(1)$$

$f(x, y)$		x			행의 합
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$		$\frac{3}{7}$
	2	$\frac{1}{28}$			$\frac{1}{28}$
열의 합		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

$\therefore X$ 와 Y 는 통계적으로 독립이 아님

결합 확률 분포

• 통계적 독립(Statistically Independent) (3/3)

• 예제 3.21

종이팩으로 포장된 부패성 식품이 보존기간(단위: 년)이 다음과 같은 확률밀도함수를 가지는 확률변수라고 하자.

$$f(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{그 외의 경우} \end{cases}$$

X_1, X_2, X_3 가 독립적으로 추출된 3개의 포장된 음식의 보존기간을 나타낸다고 할 때, $P(X_1 < 2, 1 < X_2 < 3, X_3 > 2)$ 를 구하라.

• 풀이

- 3개의 포장단위가 독립적으로 추출되었다 했으므로 세 확률변수는 통계적으로 독립

- $f(x_1, x_2, x_3) = f(x_1)f(x_2)f(x_3) = e^{-x_1}e^{-x_2}e^{-x_3} = e^{-x_1-x_2-x_3}$

- $P(X_1 < 2, 1 < X_2 < 3, X_3 > 2)$

$$= \int_2^{\infty} \int_1^3 \int_0^2 e^{-x_1-x_2-x_3} dx_1 dx_2 dx_3 = (1 - e^{-2})(e^{-1} - e^{-3})e^{-2} \approx 0.03722$$

$$\therefore 0.03722$$

Thanks!

서진우(jinwoo@pel.sejong.ac.kr)

부록#1

• 그림 코드 (1/9)

• <그림 1> 확률질량함수도

```
1. import matplotlib.pyplot as plt
2. import numpy as np

3. # 이항분포 B(4, 0.5) 확률질량함수
4. x = np.arange(0, 5)
5. f = np.array([1, 4, 6, 4, 1]) / 16

6. fig, ax = plt.subplots(figsize=(5, 4))

7. markerline, stemlines, baseline = ax.stem(x, f,
linefmt="c--", markerfmt="ko", basefmt=" ")
8. plt.setp(markerline, markersize=6)
9. plt.setp(stemlines, linewidth=1.2)

10. # y축 분수 눈금
11. yticks = np.arange(1, 7) / 16
12. ax.set_yticks(yticks)
13. ax.set_yticklabels([f"{n}/16" for n in range(1,
7)])

14. # x축 정수 눈금
15. ax.set_xticks(x)
```

```
16. ax.set_xlim(-0.4, 4.8)
17. ax.set_ylim(0, 7 / 16)

18. # 축 이름 위치 조정
19. ax.set_ylabel("f(x)", rotation=0, labelpad=15,
fontsize=12)
20. ax.yaxis.set_label_coords(0, 1.02)
21. ax.set_xlabel("x", fontsize=12)
22. ax.xaxis.set_label_coords(1.02, 0)

23. # 위·오른쪽 테두리 제거
24. ax.spines["top"].set_visible(False)
25. ax.spines["right"].set_visible(False)

26. plt.tight_layout()
27. plt.show()
```

부록#2

• 그림 코드 (2/9)

• <그림 2> 확률히스토그램

```
1. import matplotlib.pyplot as plt
2. import numpy as np

3. # 이항분포 B(4, 0.5) 확률질량함수
4. x = np.arange(0, 5)
5. f = np.array([1, 4, 6, 4, 1]) / 16

6. fig, ax = plt.subplots(figsize=(5, 4))

7. ax.bar(x, f, width=1.0, align="center",
        color="lightblue", edgecolor="black")

8. # y축 분수 눈금
9. yticks = np.arange(1, 7) / 16
10. ax.set_yticks(yticks)
11. ax.set_yticklabels([f"{n}/16" for n in range(1, 7)])

12. # x축 정수 눈금
13. ax.set_xticks(x)
```

```
14. # 축 이름 위치 조정
15. ax.set_ylabel("f(x)", rotation=0, labelpad=15,
                fontsize=12)
16. ax.yaxis.set_label_coords(0, 1.02)
17. ax.set_xlabel("x", fontsize=12)
18. ax.xaxis.set_label_coords(1.02, 0)

19. # 위·오른쪽 테두리 제거
20. ax.spines["top"].set_visible(False)
21. ax.spines["right"].set_visible(False)

22. plt.tight_layout()
23. plt.show()
```

부록#3

• 그림 코드 (3/9)

• <그림 3> $P(a < X < b)$ (1/2)

```
1. import matplotlib.pyplot as plt
2. import numpy as np

3. mu, sigma = 3.5, 1.0
4. x = np.linspace(0, 7, 400)
5. y = np.exp(-(x - mu)**2 / (2 * sigma**2))
6. a, b = mu - 0.8, mu + 1.4
7. fig, ax = plt.subplots(figsize=(6, 4))

8. # 곡선
9. ax.plot(x, y, color="steelblue", linewidth=1.3)

10. # a~b 구간 색칠
11. xs = np.linspace(a, b, 200)
12. ys = np.exp(-(xs - mu)**2 / (2 * sigma**2))
13. ax.fill_between(xs, 0, ys, color="lightblue",
alpha=0.6)
14. ax.plot([a, a], [0, np.exp(-(a - mu)**2 / (2 *
sigma**2))], color="steelblue", linewidth=1)
15. ax.plot([b, b], [0, np.exp(-(b - mu)**2 / (2 *
sigma**2))], color="steelblue", linewidth=1)
```

```
14. # 좌표축 범위
15. ax.set_xlim(-0.3, 7.6)
16. ax.set_ylim(-0.05, 1.22)

17. # 테두리 제거 (화살표 축으로 대체)
18. for spine in ax.spines.values():
19.     spine.set_visible(False)
20. ax.set_xticks([])
21. ax.set_yticks([])

22. ax.annotate("", xy=(7.6, 0), xytext=(-0.3, 0),
23.             arrowprops=dict(arrowstyle="->",
color="black", linewidth=1))
24. ax.annotate("", xy=(0, 1.22), xytext=(0, -0.05),
25.             arrowprops=dict(arrowstyle="->",
color="black", linewidth=1))

26. # 축 이름
27. ax.text(-0.08, 1.2, "f(x)", fontsize=12,
style="italic", ha="right", va="top")
28. ax.text(7.65, 0, "x", fontsize=12,
style="italic", ha="left", va="center")
```

부록#4

- 그림 코드 (4/9)
 - <그림 3> $P(a < X < b)$ (2/2)

```
29. # a, b 눈금
30. for v, label in [(a, "a"), (b, "b")]:
31.     ax.plot([v, v], [-0.02, 0.02],
32.             color="black", linewidth=1)
32.     ax.text(v, -0.09, label, fontsize=11,
33.             style="italic", ha="center", va="top")
33. plt.tight_layout()
34. plt.show()
```

부록#5

• 그림 코드 (5/9)

• <그림 4> 전형적인 밀도함수

```
1. import matplotlib.pyplot as plt
2. import numpy as np

3. fig, axs = plt.subplots(1, 4, figsize=(8, 4))

4. # (a) 오른쪽 꼬리 비대칭 분포
5. x = np.linspace(0, 6, 400)
6. y = x * np.exp(-x)
7. axs[0].plot(x, y, color="c", linewidth=1.5)
8. axs[0].set_xlim(-0.4, 6.4)
9. axs[0].set_ylim(-0.02, 0.42)

10. # (b) 증가하는 직선
11. x = np.array([0.12, 0.85])
12. y = np.array([0.22, 0.55])
13. axs[1].plot(x, y, color="c", linewidth=1.5)
14. axs[1].set_xlim(0, 1)
15. axs[1].set_ylim(0, 1)
```

```
16. # (c) 종 모양 분포
17. x = np.linspace(-3.5, 3.5, 400)
18. y = np.exp(-x**2 / 2)
19. axs[2].plot(x, y, color="c", linewidth=1.5)
20. axs[2].set_xlim(-3.9, 3.9)
21. axs[2].set_ylim(-0.03, 1.1)

22. # (d) U자형 분포
23. x = np.linspace(-1, 1, 400)
24. y = x**2
25. axs[3].plot(x, y, color="c", linewidth=1.5)
26. axs[3].set_xlim(-1.15, 1.15)
27. axs[3].set_ylim(-0.05, 1.08)

28. labels = ["(a)", "(b)", "(c)", "(d)"]
29. for ax, label in zip(axs, labels):
30.     ax.set_xticks([])
31.     ax.set_yticks([])
32.     ax.set_xlabel(label, fontsize=12)

33. plt.tight_layout()
34. plt.show()
```

부록#6

• 그림 코드 (6/9)

• <그림 5> 확률밀도함수 (1/2)

```
1. import matplotlib.pyplot as plt
2. import numpy as np

3. fig, ax = plt.subplots(figsize=(6, 4.5))

4. # 확률밀도함수  $f(x) = x^2/3$  ( $-1 < x < 2$ )
5. x_pdf = np.linspace(-1, 2, 400)
6. y_pdf = x_pdf**2 / 3

7. # 테두리 제거 (화살표 축으로 대체)
8. for spine in ax.spines.values():
9.     spine.set_visible(False)
10. ax.set_xticks([])
11. ax.set_yticks([])

12. ax.set_xlim(-2.0, 3.0)
13. ax.set_ylim(-0.18, 1.65)
```

```
14. # 화살표 축
15. ax.annotate("", xy=(2.95, 0), xytext=(-1.9, 0),
16.             arrowprops=dict(arrowstyle="->",
17.                             color="black", linewidth=1))
18. ax.annotate("", xy=(0, 1.6), xytext=(0, -0.15),
19.             arrowprops=dict(arrowstyle="->",
20.                             color="black", linewidth=1))

19. # 정의역 밖 ( $f=0$ )
20. ax.plot([-1.8, -1], [0, 0], color="steelblue",
21.         linewidth=2)
21. ax.plot([2, 2.8], [0, 0], color="steelblue",
22.         linewidth=2)

22. # 경계 수직 점선
23. ax.plot([-1, -1], [0, 1/3], color="steelblue",
24.         linewidth=1.2, linestyle="--")
24. ax.plot([2, 2], [0, 4/3], color="steelblue",
25.         linewidth=1.2, linestyle="--")

25. #  $P(0 < X \leq 1)$  구간 색칠
26. x_shade = np.linspace(0, 1, 200)
27. y_shade = x_shade**2 / 3
28. ax.fill_between(x_shade, 0, y_shade,
29.                 color="lightblue", alpha=0.7)
```

부록#7

• 그림 코드 (7/9)

• <그림 5> 확률밀도함수 (2/2)

```
29. # 포물선 곡선
30. ax.plot(x_pdf, y_pdf, color="steelblue",
           linewidth=2)

31. # 경계의 열린 원
32. for px, py in [(-1, 0), (-1, 1/3), (2, 0), (2,
4/3)]:
33.     ax.plot(px, py, "o", color="steelblue",
           markersize=6,
34.             markerfacecolor="white",
           markeredgewidth=1.5)

35. # x축 눈금
36. for v, label in [(-1, r"$-1$"), (1, r"$1$"), (2,
r"$2$")]:
37.     ax.plot([v, v], [-0.03, 0.03],
           color="black", linewidth=1)
38.     ax.text(v, -0.1, label, fontsize=11,
           ha="center", va="top")
```

```
39. # y축 눈금
40. ax.plot([-0.04, 0.04], [1/3, 1/3],
           color="black", linewidth=1)
41. ax.text(-0.08, 1/3, r"$\frac{1}{3}$",
           fontsize=12, ha="right", va="center")

42. ax.plot([-0.04, 0.04], [4/3, 4/3],
           color="black", linewidth=1)
43. ax.text(-0.08, 4/3, r"$\frac{4}{3}$",
           fontsize=12, ha="right", va="center")

44. # 함수식 표시
45. ax.text(1.55, 0.95, r"$f(x)=\frac{x^2}{3}$",
           fontsize=12, color="steelblue",
46.         ha="left", va="bottom")

47. # 축 이름
48. ax.text(-0.1, 1.58, r"$f(x)$", fontsize=12,
           style="italic", ha="right", va="top")
49. ax.text(3.0, -0.1, r"$x$", fontsize=12,
           style="italic", ha="left", va="top")

50. plt.tight_layout()
51. plt.show()
```

부록#8

• 그림 코드 (8/9)

• <그림 6> 이산형 누적분포

```
1. import matplotlib.pyplot as plt
2. import numpy as np

3. # 이항분포 B(4, 0.5) 누적분포함수
4. x_vals = np.array([0, 1, 2, 3, 4])
5. pmf = np.array([1, 4, 6, 4, 1]) / 16
6. F = np.cumsum(pmf) # [1/16, 5/16, 11/16, 15/16, 1]

7. fig, ax = plt.subplots(figsize=(5, 4))

8. # x<0 구간 (F=0)
9. ax.hlines(0, -0.5, 0, color="c", linewidth=1.5)

10. # 계단 수평선과 점
11. for i, xv in enumerate(x_vals):
12.     x_next = x_vals[i + 1] if i + 1
        < len(x_vals) else xv + 1
13.     ax.hlines(F[i], xv, x_next, color="c",
        linewidth=1.5)
14.     ax.plot(xv, F[i], "ko", markersize=6)
```

```
15. # 계단 점프 구간 점선
16. prev = np.concatenate(([0], F[:-1]))
17. for xv, y_prev, y_now in zip(x_vals, prev, F):
18.     ax.vlines(xv, y_prev, y_now, color="c",
        linestyle="--", linewidth=1.2)

19. # y축 분수 눈금
20. ax.set_yticks([0.25, 0.5, 0.75, 1])
21. ax.set_yticklabels(["1/4", "1/2", "3/4", "1"])

22. # x축 정수 눈금
23. ax.set_xticks(x_vals)

24. ax.set_xlim(-0.5, 5.3)
25. ax.set_ylim(0, 1.15)

26. # 축 이름 위치 조정
27. ax.set_ylabel("F(x)", rotation=0, labelpad=15,
        fontsize=12, style="italic")
28. ax.yaxis.set_label_coords(0, 1.05)
29. ax.set_xlabel("x", fontsize=12, style="italic")
30. ax.xaxis.set_label_coords(1.02, 0)

31. # 위·오른쪽 테두리 제거
32. ax.spines["top"].set_visible(False)
33. ax.spines["right"].set_visible(False)

34. plt.tight_layout()
35. plt.show()
```

부록#9

• 그림 코드 (9/9)

• <그림 7> 연속형 누적분포

```
1. import numpy as np
2. import matplotlib.pyplot as plt

3. x = np.linspace(-1, 2.6, 800)
4. y = np.zeros_like(x)

5. # -1<=x<0: 낮은 값까지 빠르게 상승
6. mask1 = (x >= -1) & (x < 0)
7. y[mask1] = 0.12 * (1 - np.exp(-4 * (x[mask1] + 1)))

8. # 0<=x<2: 지수 형태로 상승
9. mask2 = (x >= 0) & (x < 2)
10. y0 = 0.12 * (1 - np.exp(-4))
11. A = (1 - y0) / (np.exp(4) - 1)
12. y[mask2] = y0 + A * (np.exp(2 * x[mask2]) - 1)

13. # x>=2: 1.0에서 평평
14. mask3 = x >= 2
15. y[mask3] = 1.0

16. fig, ax = plt.subplots(figsize=(5.2, 3.1))
```

```
17. ax.plot(x, y, color="deepskyblue",
           linewidth=1.5)

18. # 축을 원점으로 이동, 위·오른쪽 테두리 제거
19. ax.spines["left"].set_position("zero")
20. ax.spines["bottom"].set_position("zero")
21. ax.spines["top"].set_visible(False)
22. ax.spines["right"].set_visible(False)

23. ax.set_xlim(-1.2, 2.6)
24. ax.set_ylim(-0.05, 1.15)

25. ax.set_xticks([-1, 0, 1, 2])
26. ax.set_yticks([0.5, 1.0])
27. ax.set_yticklabels(["0.5", "1.0"])

28. # 축 이름
29. ax.text(2.63, 0, "x", fontsize=12,
          style="italic", ha="left", va="center")
30. ax.text(0, 1.2, "f(x)", fontsize=12,
          style="italic", ha="center", va="bottom")

31. plt.tight_layout()
32. plt.show()
```