2024/01/03, 확률 기초 세미나

확률 및 통계학

-3장 확률변수와 확률분포-

이 하 늘(haneul@pel.sejong.ac.kr)

세종대학교 프로토콜공학연구실

목 차

- 확률변수의 개념
- 이산형 확률분포
- 연속형 확률분포
- 결합확률분포

목 차

확률변수의 개념

- 확률변수(Random Variable)

정의 3.1

확률변수는 표본공간 내의 각 원소에 하나의 실수 값을 대응시키는 함수로 정의된다.

  - 특징
    - 확률변수는 주로 대문자로 나타내고, 그에 대응되는 하나의 값은 소문자로 나타냄
      - e.g., 확률변수 $X \rightarrow$ 대응되는 값 x
    - 확률변수 X의 모든 값들은 표본공간의 부분집합이 되는 사상을 나타냄
    - 확률변수에는 이산형 확률변수와 연속형 확률변수가 존재함
      - 이산형 확률변수(Discrete Random Variable)
        - 셀 수 있는 유한한 값을 가지는 확률변수
      - 연속형 확률변수(Continuous Random Variable)
        - 셀 수 없는 무한한 값을 가지는 확률변수

확률변수의 개념

- 예제 3.1

4개의 붉은 공(R)과 3개의 검은 공(B)이 들어 있는 항아리에서 **연속적으로 2개의 공을 비복원추출**하는 실험에서 **Y를 붉은 공의 개수**라 할 때, 출현가능한 결과와 확률변수 Y의 값 y를 구하라.

- $S = \{\mathrm{RR}, \mathrm{RB}, \mathrm{BR}, \mathrm{BB}\}$
- $Y = \{0, 1, 2\}$

표본공간	y
RR	2
RB	1
BR	1
BB	0

확률변수의 개념

- 예제 3.2

공구보관소의 직원이 **3명의 공장 종업원**들에게 안전헬멧을 임의로 꺼내 주었을 경우, **스미스(S), 존스(J), 그리고 브라운(B)의 순서**로 헬멧을 받을 때, 헬멧을 받는 가능한 순서들을 나열하고, **M을 헬멧이 원래 주인에게 지급되는 경우의 수**라 할 때, 확률변수 M의 값 m을 구하라.

- $S = \{\text{SJB, SBJ, JSB, JBS, BSJ, BJS}\}$
- $M = \{0, 1, 3\}$

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

확률변수의 개념

- 가변수(Dummy Variable)
  - 정의
    - 객관적인 통계 수치로 나타내기 어려운 범주형 변수를 이산형 혹은 연속형 변수로 변환하여 나타내는 변수
      - 범주형 변수: 실험 결과가 숫자가 아닌 몇 개의 범주 혹은 항목으로 나타나는 정성적 변수

성별
여성
남성
남성
여성
여성
남성
…

성별
1
0
0
1
1
0
…

<가변수 예시 1>

흡연여부
비흡연
흡연
비흡연
비흡연
흡연
비흡연
…

흡연여부
0
1
0
0
1
0
…

<가변수 예시 2>

확률변수의 개념

- 이산표본공간(Discrete Sample Space)

정의 3.2

표본공간이 유한개 혹은 셀 수 있는 무한개 원소로 이루어졌을 때 **이산표본공간(Discrete Sample Space)**이라 한다.

  - 계수자료(Count Data)를 나타냄
    - e.g., 불량품 개수, 사고횟수 등

  - 예제 3.3

공장에서 생산되는 부품들을 불량이나 양품으로 판정한다고 하고 확률변수 X를 정의하라.

- $X = \begin{cases} 1, \text{부품이 불량일 때} \\ 0, \text{부품이 양품일 때} \end{cases}$
- 두 개의 가능한 값을 0과 1로 표현하는 확률변수를 **베르누이 확률변수(Bernoulli random variable)**라고 함

확률변수의 개념

- 이산표본공간(Discrete Sample Space)
  - 예제 3.4

생산품의 합격 또는 불합격 판정을 위해, 12개의 불량품이 있는 100개의 제품에서 **10개를 독립적으로 추출**하는 샘플링 검사법(Sampling plan)을 사용하는 경우, 10개 제품 표본에서 발견되는 불량품의 수를 확률변수 X일 때, X가 가질 수 있는 값을 구하라.

- $X = \{0, 1, 2, \cdots, 9, 10\}$

  - 예제 3.5

하나의 불량품이 발견될 때까지 공정으로부터 표본을 추출하는 샘플링 검사법에서, **불량품이 발견될 때까지 추출한 제품의 수를 나타내는 확률변수를 X**라고 하자. 양품을 N, 불량품을 D로 나타낼 때, 표본공간을 나타내라.

- $X = 1$이면 $S = \{\mathrm{D}\}$, $X = 2$이면 $S = \{ND\}$, $X = 3$이면 $S = \{NND\}$, $\cdots$

확률변수의 개념

- 연속표본공간(Continuous Sample Space)

정의 3.3

표본공간이 실선의 어떤 구간 내의 모든 수를 포함할 때 **연속표본공간(Continuous Sample Space)**이라 한다.

  - 측정자료(Measured Data)를 나타냄
    - e.g., 높이, 무게, 온도, 거리, 수명 등
  - 예제 3.6

통신판매 광고에 반응하는 사람들의 **비율**을 X라고 하면, 확률변수 X의 값 x의 범위를 구하라.

- $0 \leq x \leq 1$

  - 예제 3.7

과속탐지 카메라에 적발되는 과속 차량들 사이의 시간간격을 확률변수X라고 하면, X의 값 x의 범위를 구하라.

- $x \geq 0$

목 차

- 확률변수의 개념
- 이산형 확률분포
- 연속형 확률분포
- 결합확률분포

이산형 확률분포

- 확률분포(Probability Distribution)
  - 정의
    - 표본공간에 정의된 확률을 통해 확률변수의 값 혹은 확률변수의 집합에 대한 확률을 표현한 것
  - 이산형 확률분포(Discrete Probability Distribution)

정의 3.4

모든 x에 대해 순서쌍 $(x, f(x))$의 집합이 다음 조건을 만족하면 이를 **이산형 확률변수 X의 확률함수, 확률질량함수(PMF, Probability Mass Function), 혹은 확률분포**라고 한다.

1. $f(x) \geq 0$
2. $\sum_x f(x) = 1$
3. $P(X = x) = f(x)$

이산형 확률분포

- 예제 3.8

상점에 진열된 **20대의 노트북 중에 불량품이 3대 포함**되어 있는 경우, 어느 학교에서 **이 중 임의로 2대를 구입**했을 때, 불량품 개수의 확률분포를 구하라.

- 학교에서 구입한 노트북 중 불량품의 수: 확률변수 X
- X의 값 x는 0, 1, 2 중에서 값을 취할 수 있음에 따라, 확률분포는 다음과 같이 구할 수 있음

$$f(0) = P(X=0) = \frac{\binom{3}{0}\binom{17}{2}}{\binom{20}{2}} = \frac{68}{95}, \; f(1) = P(X=1) = \frac{\binom{3}{1}\binom{17}{1}}{\binom{20}{2}} = \frac{51}{190}$$

$$f(2) = P(X=2) = \frac{\binom{3}{2}\binom{17}{0}}{\binom{20}{2}} = \frac{3}{190}$$

- X의 확률분포는 다음과 같이 정리할 수 있음

$$f(x) = \begin{cases} \frac{68}{95}, & x = 0 \\ \frac{51}{190}, & x = 1 \\ \frac{3}{190}, & x = 2 \end{cases}$$

이산형 확률분포

- 예제 3.9

어느 대리점에서 **판매된 외제차의 50%에 디젤엔진이 장착**되었다고 할 때, 이 대리점에서 다음에 판매될 **4대의 외제차 가운데 디젤엔진이 장착된 차의 수의 확률분포에 대한 식**을 구하라.

- 디젤엔진이 장착된 확률 : $\frac{1}{2}$, 가솔린엔진이 장착된 확률 : $\frac{1}{2}$
- 표본공간에서 표본점의 개수 : $2^4 = 16$ 개

- 4대의 외제차 중에 디젤엔진이 장착된 경우가 x라면, 4대의 외제차 중에 가솔린엔진이 장착된 경우는 $4 - x$
- 4대의 외제차 중에 디젤모델을 고르는 경우의 수는 $\binom{4}{x}$로 표현할 수 있으며, 따라서 확률분포에 대한 식은 다음과 같이 구할 수 있음

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

이산형 확률분포

- 누적분포(Cumulative Distribution)
  - 확률변수 X의 값이 어떤 실수 x보다 작거나 같은 확률을 계산해야 하는 경우, 모든 실수 x에 대한 $F(x) = P(X \leq x)$를 확률변수 X의 누적분포라고 함
  - 누적분포는 확률적인 추론 및 결정 등에 활용할 수 있음

정의 3.5

확률분포 $f(x)$를 가지는 이산형 확률변수 X의 **누적분포함수(CDF, Cumulative Distribution Function)** $F(x)$는

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), -\infty < x < \infty$$

로 주어진다.

이산형 확률분포

- 예제 3.10

예제 3.9에서 확률변수 X의 누적분포를 구하라. 그리고 $F(x)$를 사용하여 $f(2)=\frac{3}{8}$이 됨을 증명하라.

- 디젤엔진 외제차를 고르는 확률변수 X의 확률분포
  - $f(0)=\frac{1}{16}, f(1)=\frac{1}{4}, f(2)=\frac{3}{8}, f(3)=\frac{1}{4}, f(4)=\frac{1}{16}$

- 누적분포를 구하면

$$F(0)=f(0)=\frac{1}{16}$$
$$F(1)=f(0)+f(1)=\frac{5}{16}$$
$$F(2)=f(0)+f(1)+f(2)=\frac{11}{16}$$
$$F(3)=f(0)+f(1)+f(2)+f(3)=\frac{15}{16}$$
$$F(4)=f(0)+f(1)+f(2)+f(3)+f(4)=1$$

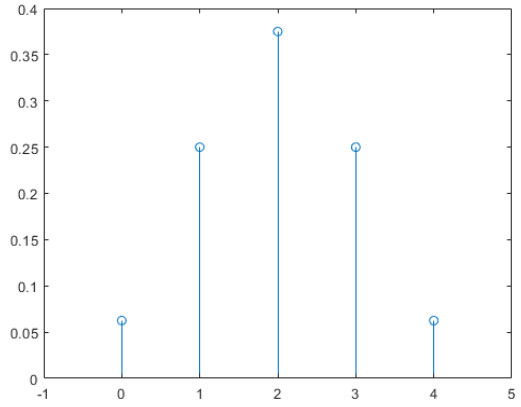
$$F(x)=\begin{cases}0, & x<0\\ \frac{1}{16}, & 0\le x<1\\ \frac{5}{16}, & 1\le x<2\\ \frac{11}{16}, & 2\le x<3\\ \frac{15}{16}, & 3\le x<4\\ 1, & 4\le x\end{cases}$$

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

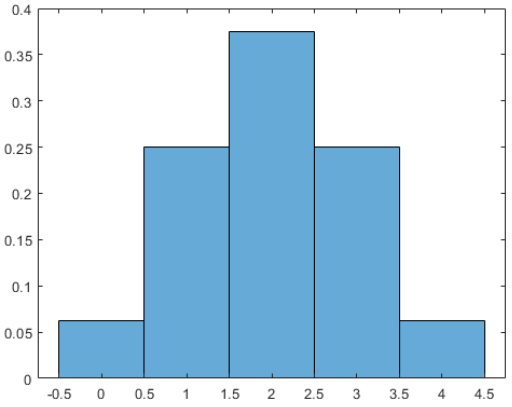
이산형 확률분포

- 그래프 종류

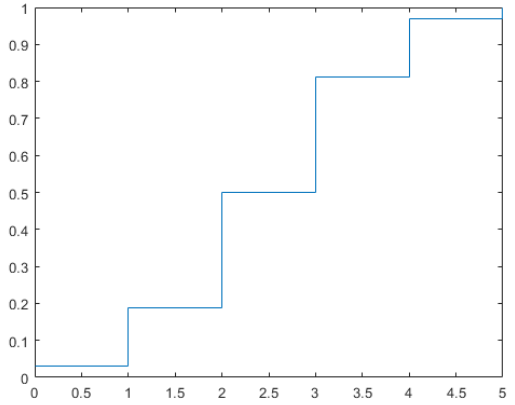
1. 확률질량함수도(Probability Mass Function Plot)
   - 확률 변수의 값들을 가시적으로 표현함으로써, 확률분포를 쉽게 파악 가능
2. 확률히스토그램(Probability Histogram)
   - 막대 면적을 이용하여 특정 구간에서 발생할 확률을 추정 가능
3. 이산형 누적분포(Discrete Cumulative Distribution)
   - 확률 변수가 특정 값 이하일 확률을 직관적으로 파악 가능



<그림 3.1> 확률질량함수도



<그림 3.2> 확률히스토그램



<그림 3.3> 이산형 누적분포

목 차

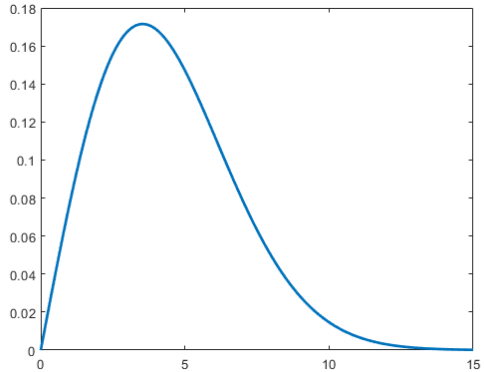
연속형 확률분포

- 연속형 확률분포(Continuous Probability Distribution)
  - 연속형 확률변수와 이에 대응하는 확률을 표현한 것

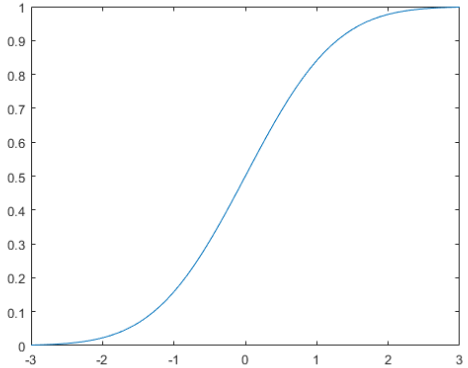
정의 3.6

다음 조건이 만족되면 $f(x)$를 실수의 집합 R 상에서 정의된 연속형 확률변수에 대한 **확률밀도함수(PDF, Probability Density Function)**라고 한다.

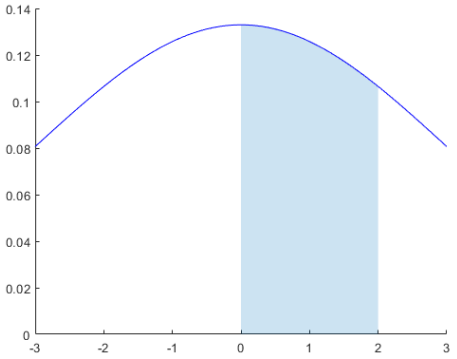
1. 모든 $x \in R$에 대하여 $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
3. $P(a < X < b) = \int_{a}^{b} f(x)\,dx$



<그림 3.4(1)> 확률밀도함수 예시1



<그림 3.4(2)> 확률밀도함수 예시2



<그림 3.5> $P(0 < X < 2)$ 예시

연속형 확률분포

- 예제 3.11

제어실험에서 반응온도(˚C)변화에 따른 오차는 다음과 같은 확률분포를 가지는 연속확률변수 X라고 가정하자.

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

- (a) $f(x)$가 확률밀도함수임을 증명하라.
  - $f(x) \geq 0$임은 명확함에 따라, 다음과 같이 조건을 확인하여 증명할 수 있다.

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

- (b) $P(0 < X \leq 1)$을 구하라.

$$\mathrm{P}(0 < X \leq 1) = \int_{0}^{1} \frac{x^2}{3} dx = \frac{x^3}{9}\Big|_{0}^{1} = \frac{1}{9}$$

연속형 확률분포

- 누적분포함수(CDF, Cumulative Distribution Function)

정의 3.7

확률밀도함수가 $f(x)$인 연속형 확률변수 X의 **누적분포함수** $F(x)$는 다음과 같다.

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt, \qquad -\infty < x < \infty$$

- 정의 3.7에 의해 알 수 있는 사실
  - $P(a < X < b) = F(b) - F(a)$
  - 누적분포함수가 미분이 가능하면, $f(x) = \frac{dF(x)}{dx}$

연속형 확률분포

- 예제 3.12

예제 3.11의 확률밀도함수에 대하여 **F(x)를 구하고, 그것을 이용하여 P(0 < X ≤ 1)을 구하라**.

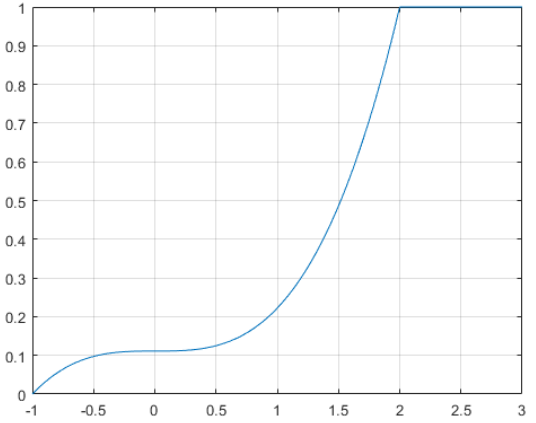
$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

- $-1 < x < 2$에 대하여

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

따라서,

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



<그림 3.6> 연속형 누적분포함수

- $P(0 < X \le 1)$를 구하게 되면 다음과 같이 구할 수 있으므로, 예제 3.11의 결과와 같다.

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

연속형 확률분포

- 예제 3.13

프로젝트를 입찰에 부치고 입찰가를 예상할 때, **예상치를 b**라고 하면, **낙찰가 y**에 대한 밀도함수는 다음과 같다.

$$f(y) = \begin{cases} \frac{5}{8b}, & \frac{2}{5}b \le y \le 2b \\ 0, & \text{otherwise} \end{cases}$$

$F(y)$를 구하고, 낙찰가가 예상치 b보다 작을 확률을 구하라.

- $\frac{2}{5}b \le y \le 2b$에 대하여,

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

이다. 따라서,

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

낙찰가가 예상치보다 작을 확률은 다음과 같다.

$$P(Y \le b) = F(b) = \frac{5}{8} - \frac{1}{4} = \frac{3}{8}$$

목 차

결합확률분포

- 결합확률분포(Joint Probability Distribution)
  - 여러 개의 확률변수들의 결과를 동시에 취급하는 경우, 두 확률변수 X, Y의 값에 대응하는 확률을 표현한 것

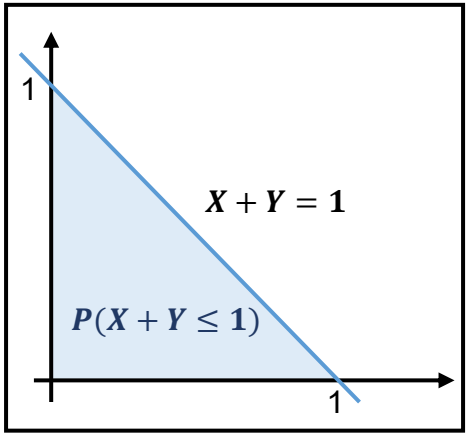
- 결합확률질량함수(Joint PMF, Joint Probability Mass Function)

정의 3.8

다음 조건이 만족될 때 함수 $f(x,y)$를 **이산형 확률변수 X와 Y의 결합확률분포** 또는 **결합확률질량함수(Joint PMF)**라 한다.

1. 모든 (x,y)에 대하여 $f(x,y) \geq 0$
2. $\sum_x \sum_y f(x,y) = 1$
3. $P(X = x, Y = y) = f(x,y)$

평면상의 어떤 영역 A에 대하여 $P[(X,Y) \in A] = \sum\sum_A f(x,y)$ 가 된다.



<그림 3.7> 영역 $X+Y \leq 1$ 예시

결합확률분포

- 예제 3.14

3개의 청색, 2개의 적색, 3개의 녹색 볼펜이 들어 있는 상자에서 임의로 **2개를 추출**하고자 하는 경우, ***X*를 청색 볼펜의 수, *Y*를 적색 볼펜의 수**라고 하자.

- (a) 결합확률분포 $f(x,y)$를 구하라.
  - 8개 중 임의로 2개의 볼펜을 뽑는 경우의 수: $\binom{8}{2}$
  - 결합확률질량함수는 다음과 같다.

$$f(x,y)=\frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{\binom{8}{2}}, \qquad \mathrm{x}=0,1,2,\mathrm{y}=0,1,2,0\le \mathrm{x}+\mathrm{y}\le 2$$

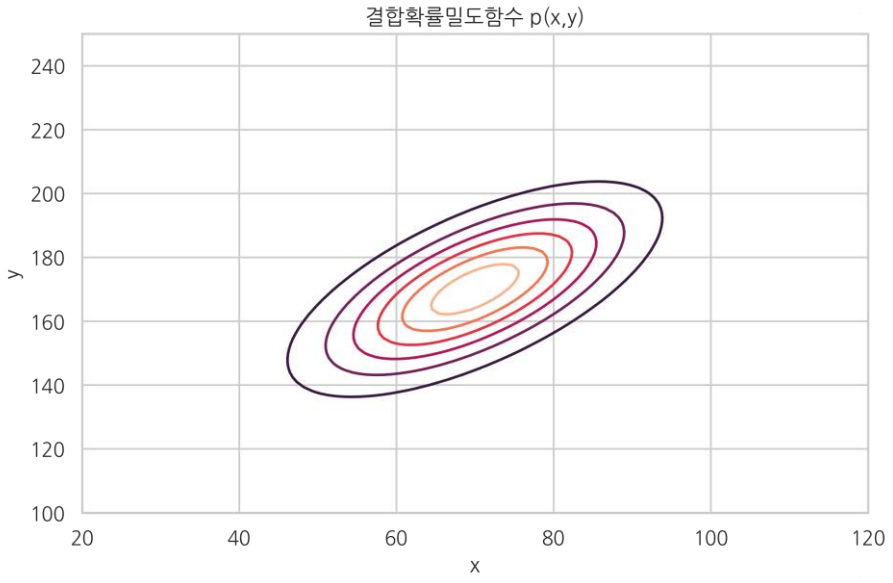
$f(x,y)$		x			행의 합
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
열의 합		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

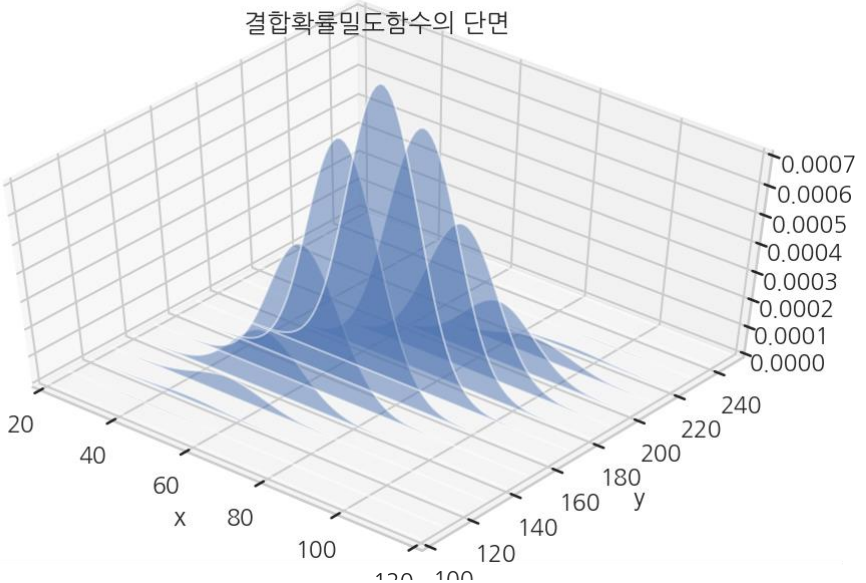
<표 3.1> 예제3.14의 결합확률분포 표

- (b) $A=\{(x,y)|x+y\le 1\}$이라고 할 때, $P[(X,Y)\in A]$를 구하라.
  - $P[(X,Y)\in A]=P(X+Y\le 1)=f(0,0)+f(0,1)+f(1,0)=\frac{3}{28}+\frac{3}{14}+\frac{9}{28}=\frac{9}{14}$

결합확률분포

- 결합확률밀도함수(joint PDF, joint Probability Density Function)
  - X와 Y가 연속형 확률변수이면 $f(x,y)$는 xy평면 위에 놓여 있는 표면이 됨
  - A를 xy평면상의 임의의 영역이라면 $P(X,Y) \in A$는 밑면과 단면으로 구성되는 입체의 부피와 같게 됨





<출처>: https://datascienceschool.net

결합확률분포

- 결합확률밀도함수(joint PDF, joint Probability Density Function)

정의 3.9

다음 조건이 만족될 때 함수 $f(x,y)$를 연속확률변수 X와 Y의 **결합확률밀도함수(joint PDF)**라 한다.

1. 모든 (x,y)에 대하여 $f(x,y) \geq 0$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$
3. $P[(X,Y) \in A] = \int\int_{A} f(x,y)\,dx\,dy$, 이때 A는 xy 평면상의 임의의 영역

결합확률분포

- 예제 3.15 (1/2)

어느 과자회사에서는 연한 초콜릿과 진한 초콜릿을 입힌 과자상자를 취급할 때, 임의로 하나의 과자상자를 선택하는 경우, **X와 Y를 각각 연한 초콜릿과 진한 초콜릿의 비율**이라 하면, 결합확률 분포는 다음과 같다.

$$f(x,y)=\begin{cases}\frac{2}{5}(2x+3y), & 0\le x\le 1, 0\le y\le 1\\ 0, & \text{otherwise}\end{cases}$$

- (a) 정의 3.9의 조건 2를 증명하라.
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_0^1\int_0^1 \frac{2}{5}(2x+3y)\,dx\,dy$

$$= \int_0^1 \left(\frac{2x^2}{5}+\frac{6xy}{5}\right)\Big|_{x=0}^{x=1} dy$$

$$= \int_0^1 \left(\frac{2}{5}+\frac{6y}{5}\right) dy = \left(\frac{2y}{5}+\frac{3y^2}{5}\right)\Big|_0^1 = \frac{2}{5}+\frac{3}{5}=1$$

결합확률분포

- 예제 3.15 (2/2)

어느 과자회사에서는 연한 초콜릿과 진한 초콜릿을 입힌 과자상자를 취급할 때, 임의로 하나의 과자상자를 선택하는 경우, **X와 Y를 각각 연한 초콜릿과 진한 초콜릿의 비율**이라 하면, 결합확률 분포는 다음과 같다.

$$f(x,y)=\begin{cases}\frac{2}{5}(2x+3y), & 0\le x\le 1, 0\le y\le 1\\ 0, & \text{otherwise}\end{cases}$$

- (b) $A=\{(x,y)|0<x<\frac{1}{2},\ \frac{1}{4}<y<\frac{1}{2}\}$이라고 할 때, $P[(X,Y)\in A]$를 구하라.
  - $P[(X,Y)\in A]=\mathrm{P}\left(0<X<\frac{1}{2},\frac{1}{4}<Y<\frac{1}{2}\right)$

$$=\int_{\frac{1}{4}}^{\frac{1}{2}}\int_{0}^{\frac{1}{2}}\frac{2}{5}(2x+3y)\,dx\,dy=\int_{\frac{1}{4}}^{\frac{1}{2}}\left(\frac{2x^2}{5}+\frac{6xy}{5}\right)\Big|_{x=0}^{x=\frac{1}{2}}dy$$

$$=\int_{\frac{1}{4}}^{\frac{1}{2}}\left(\frac{1}{10}+\frac{3y}{5}\right)dy=\left(\frac{y}{10}+\frac{3y^2}{10}\right)\Big|_{\frac{1}{4}}^{\frac{1}{2}}$$

$$=\frac{1}{10}\left[\left(\frac{1}{2}+\frac{3}{4}\right)-\left(\frac{1}{4}+\frac{3}{16}\right)\right]=\frac{13}{160}$$

결합확률분포

- 주변분포(Marginal Distribution)
  - 결합분포에서 하나의 확률변수에 대한 확률분포를 추출하는 것
  - 다변량 데이터에 대한 개별 데이터의 특성 파악, 조건부 분포 계산 등을 위해 사용됨

정의 3.10

X와 Y의 **주변분포**는 다음과 같이 주어진다.

1. 이산형인 경우

$$g(x) = \sum_{y} f(x,y), \qquad h(y) = \sum_{x} f(x,y)$$

2. 연속형인 경우

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

결합확률분포

- 예제 3.16

표 3.1의 열과 행의 합이 각각 X와 Y의 주변분포가 됨을 증명하라.

- 확률변수 X에 대하여 다음 값들을 구할 수 있다.
  - $g(0) = f(0,0) + f(0,1) + f(0,2) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14}$
  - $g(1) = f(1,0) + f(1,1) + f(1,2) = \frac{9}{28} + \frac{3}{14} + 0 = \frac{15}{28}$
  - $g(2) = f(2,0) + f(2,1) + f(2,2) = \frac{3}{28} + 0 + 0 = \frac{3}{28}$

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

<표 3.1> 예제3.14의 결합확률분포 표

- 위의 값들은 표3.1의 열의 합과 일치하며, 같은 방법으로 $h(y)$의 값들이 행의 합으로 표현됨을 알 수 있다. 이에 대한 주변분포는 다음과 같이 나타낼 수 있다.

$$g(x) = \begin{cases} \frac{5}{14}, & x = 0 \\ \frac{15}{28}, & x = 1 \\ \frac{3}{28}, & x = 2 \end{cases} \qquad h(\mathrm{y}) = \begin{cases} \frac{15}{28}, & \mathrm{y} = 0 \\ \frac{3}{7}, & \mathrm{y} = 1 \\ \frac{1}{28}, & \mathrm{y} = 2 \end{cases}$$

결합확률분포

- 예제 3.17

예제 3.15의 결합분포에 대하여 $g(x)$와 $h(y)$를 구하라.

$$f(x,y)=\begin{cases}\frac{2}{5}(2x+3y), & 0\le x\le 1, 0\le y\le 1\\ 0, & \text{otherwise}\end{cases}$$

- $0\le x\le 1$ 구간에서는 다음과 같으며, 그 외의 영역에서는 $g(x)=0$이 된다.

$$g(x)=\int_{-\infty}^{\infty} f(x,y)\,dy=\int_0^1 \frac{2}{5}(2x+3y)\,dy=\left(\frac{4xy}{5}+\frac{3y^2}{5}\right)\Bigg|_{y=0}^{y=1}=\frac{4x+3}{5}$$

- 같은 방법으로 $0\le y\le 1$ 구간에서는 다음과 같으며, 그 외의 영역에서는 $h(y)=0$이 된다.

$$h(y)=\int_{-\infty}^{\infty} f(x,y)\,dx=\int_0^1 \frac{2}{5}(2x+3y)\,dx=\frac{2(1+3y)}{5}$$

결합확률분포

- 조건부 분포(Conditional Distribution)
  - 조건이 주어진 경우, 특정 확률변수에 대한 확률분포를 나타내는 것
  - 주어진 조건에 따른 데이터의 분포를 분석하는데 사용됨
    - e.g., 사용자 행동 패턴 분석 등

정의 3.11

X와 Y를 이산형 또는 연속형인 두 확률변수라고 할 때, $X = x$로 주어졌을 때 확률변수 Y의 조건부 분포(Conditional Distribution)는 다음과 같이 주어진다.

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

같은 방법으로 $Y = y$로 주어졌을 때 확률변수 X의 조건부 분포는 다음과 같이 주어진다.

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

결합확률분포

- 조건부 분포(Conditional Distribution)
  - 확률변수가 a와 b사이의 값을 가질 확률
    - 확률변수 X와 Y가 이산형 확률변수인 경우
      - $\mathrm{P}(a < X < b | Y = y) = \sum_{a<x<b} f(x|y)$
    - 확률변수 X와 Y가 연속형 확률변수인 경우
      - $\mathrm{P}(a < X < b | Y = y) = \int_a^b f(x|y)dx$

결합확률분포

- 예제 3.18 (1/3)

예제 3.14에서 $\mathrm{Y}=1$로 주어졌을 때, X의 조건부 분포를 구하고, 그것을 이용하여 $\mathrm{P}(\mathrm{X}=0|\mathrm{Y}=1)$을 구하라.

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

<표 3.1> 예제3.14의 결합확률분포 표

$Y=1$일 때, $f(x|y)$가 필요하므로 $h(1)$은 다음과 같이 구할 수 있다.

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

$f(x|1)$은 다음과 같이 구할 수 있다.

$$f(x|1)=\frac{f(x,1)}{h(1)}=\frac{7}{3}f(x,1), \qquad x=0,1,2$$

결합확률분포

- 예제 3.18 (2/3)

예제 3.14에서 $\mathrm{Y}=1$로 주어졌을 때, X의 조건부 분포를 구하고, 그것을 이용하여 $\mathrm{P}(\mathrm{X}=0|\mathrm{Y}=1)$을 구하라.

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

<표 3.1> 예제3.14의 결합확률분포 표

따라서,

$$f(0|1)=\left(\frac{7}{3}\right)f(0,1)=\left(\frac{7}{3}\right)\left(\frac{3}{14}\right)=\frac{1}{2}$$
$$f(1|1)=\left(\frac{7}{3}\right)f(1,1)=\left(\frac{7}{3}\right)\left(\frac{3}{14}\right)=\frac{1}{2}$$
$$f(2|1)=\left(\frac{7}{3}\right)f(2,1)=\left(\frac{7}{3}\right)(0)=0$$

이 되고, $\mathrm{Y}=1$일 때, X 의 조건부 분포는 다음과 같다.

$$f(x|1)=\begin{cases}\frac{1}{2}, & x=0\\ \frac{1}{2}, & x=1\\ 0, & x=2\end{cases}$$

결합확률분포

- 예제 3.18 (3/3)

예제 3.14에서 $\mathrm{Y}=1$로 주어졌을 때, X의 조건부 분포를 구하고, 그것을 이용하여 $\mathrm{P}(\mathrm{X}=0|\mathrm{Y}=1)$을 구하라.

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

<표 3.1> 예제3.14의 결합확률분포 표

끝으로, 조건부 확률의 정의를 사용하면,

$$P(X=x|Y=y)=\frac{P(Y=y,X=x)}{P(Y=y)}=\frac{f(x,y)}{h(y)}=f(x|y)$$

이므로,

$$P(X=0|Y=1)=f(0|1)=\frac{1}{2}$$

따라서, 두 개의 볼펜 중 하나가 적색이라는 사실이 알려지면 다른 하나의 볼펜이 청색이 아닐 확률은 $\frac{1}{2}$이 됨을 알 수 있다.

결합확률분포

- 예제 3.19

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

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

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

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{x}^{1} 10xy^2\,dy$$
$$= \frac{10}{3}xy^3\Big|_{y=x}^{y=1} = \frac{10}{3}x(1-x^3), \qquad 0 < x < 1$$
$$h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_{0}^{y} 10xy^2\,dx = 5x^2y^2\Big|_{x=0}^{x=y} = 5y^4, \qquad 0 < y < 1$$

따라서,

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

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

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

결합확률분포

- 예제 3.20

파주시의 **온도 변화량이 X**, **비가 올 확률이 Y** 일 때, 결합확률분포가 다음과 같이 주어졌다. 이를 이용하여, **$g(x), h(y), f(x|y)$를 구하고, $P(1/4 < X < 1/2 | Y = 1/3)$의 값**을 구하라.

$$f(x,y) = \begin{cases} \dfrac{x(1+3y^2)}{4}, & 0 < x < 2, \quad 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 \frac{x(1+3y^2)}{4} dy = \left(\frac{xy}{4} + \frac{xy^3}{4}\right)\Big|_{y=0}^{y=1} = \frac{x}{2}, \qquad 0 < x < 2$$

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^2 \frac{x(1+3y^2)}{4} dx = \left(\frac{x^2}{8} + \frac{3x^2y^2}{8}\right)\Big|_{x=0}^{x=2} = \frac{1+3y^2}{2}, \qquad 0 < y < 1$$

따라서,

$$f(x|y) = \frac{f(x,y)}{h(y)} = \frac{x(1+3y^2)/4}{(1+3y^2)/2} = \frac{x}{2}, \quad 0 < x < 2$$

이고,

$$P\left(\frac{1}{4} < X < \frac{1}{2} \middle| Y = \frac{1}{3}\right) = \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{x}{2} dx = \frac{3}{64}$$

이 된다.

결합확률분포

- 통계적 독립(Statistically Independent)

정의 3.12

X와 Y를 결합확률분포 $f(x, y)$와 주변분포 $g(x)$, $h(y)$를 가지는 이산형 혹은 연속형 확률변수라 할 때, 모든 x, y 에 대하여 $f(x, y) = g(x)h(y)$가 성립하면 **확률변수 X와 Y는 통계적으로 독립**이라 함

- $f(x,y)$가 y에 종속되어 있지 않다면, $f(x|y) = \mathrm{g(x)}$이고 $f(x,y) = g(x)h(y)$가 됨
  - 증명
    - $f(x,y) = f(x|y)h(y)$
      $g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{-\infty}^{\infty} f(x|y)h(y)\,dy$
      $g(x) = f(x|y)\int_{-\infty}^{\infty} h(y)\,dy$
      $\int_{-\infty}^{\infty} h(y)\,dy = 1$, 따라서 $g(x) = f(x|y)$
      $f(x,y) = g(x)h(y)$

결합확률분포

- 예제 3.21

예제 3.14의 확률변수들이 통계적으로 독립이 아님을 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	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
열의 합		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

<표 3.1> 예제3.14의 결합확률분포 표

- $f(0,1) = \frac{3}{14}$
- $g(0) = \sum_{y=0}^{2} f(0,y) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14}$
- $h(1) = \sum_{x=0}^{2} f(x,1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}$

- $f(0,1) \neq g(0)h(1)$

결합확률분포

- 통계적 독립(Statistically Independent)

정의 3.13

$X_1, X_2, \cdots, X_n$을 결합확률분포 $f(x_1, x_2, \cdots, x_n)$과 주변분포 $f_1(x_1), f_2(x_2), \cdots, f_n(x_n)$을 가지는 이산형 혹은 연속형 확률변수라고 할 때, 모든 $(x_1, x_2, \cdots, x_n)$에 대하여

$$f(x_1, x_2, \cdots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n)$$

이 성립하면 $X_1, X_2, \cdots, X_n$을 **상호 통계적으로 독립**이라고 한다.

결합확률분포

- 예제 3.22

종이팩으로 포장된 부패성 식품의 보존기간이 다음과 같은 확률밀도함수를 가지는 확률변수라고 하자.

$$f(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & otherwise \end{cases}$$

X_1, X_2, X_3가 독립적으로 추출된 3개의 포장된 음식의 보존기간을 나타낸다고 할 때 **$P(X_1 < 2, 1 < X_2 < 3, X_3 > 2)$**를 구하라.

- 3개의 포장단위가 독립적으로 추출됨에 따라 확률변수 X_1, X_2, X_3가 통계적으로 독립이라 할 수 있고, 따라서 결합밀도함수는 다음과 같다.

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

따라서,

$$P(X_1 < 2, 1 < X_2 < 3, X_3 > 2) = \int_2^{\infty}\int_1^3\int_0^2 e^{-x_1-x_2-x_3}\,dx_1\,dx_2\,dx_3$$
$$= (1-\mathrm{e}^{-2})(\mathrm{e}^{-1}-\mathrm{e}^{-3})\mathrm{e}^{-2} = 0.0372$$

Thanks!

이 하 늘(haneul@pel.sejong.ac.kr)

부록 #1

- MATLAB 코드(1/4)
  - 그림 3.1

```
a = [0 1 2 3 4]; p = [1/16 4/16 6/16 4/16 1/16];
stem(a,p);
set(gca, 'xlim', [-1 5]);
```

  - 그림 3.2

```
histogram('BinEdges',-0.5:4.5,'BinCounts',[1/16 4/16 6/16 4/16 1/16]);
```

부록 #2

- MATLAB 코드(2/4)
  - 그림 3.3

```
x=0:5;
y=binocdf(x,5,0.5);
stairs(x,y)
```

  - 그림 3.4(1)

```
pd = makedist('Weibull','A',5,'B',2);

x = 0:.1:15;
y = pdf(pd,x);

plot(x,y,'LineWidth',2);
```

부록 #3

- MATLAB 코드(3/4)
  - 그림 3.4(2)

```
pd = makedist('Normal');

x = -3:.1:3;
p = cdf(pd,x);

plot(x,p);
```

  - 그림 3.5

```
x = -3:.1:3;
xs=x(x>=0&x<=2);

figure;
hold on;
a=area(xs,normpdf(xs,0,3));
a.FaceAlpha=0.2;
p=plot(x,normpdf(x,0,3));
p.Color='blue';
```

부록 #4

- MATLAB 코드(4/4)
  - 그림 3.6

```
x = linspace(-1, 3, 1000);

f1 = @(x) 0;
f2 = @(x) (x.^3 + 1) / 9;
f3 = @(x) 1;

y = zeros(size(x));
y(x < -1) = f1(x(x < 0));
y(x >= -1 & x < 2) = f2(x(x >= -1 & x < 2));
y(x >= 2) = f3(x(x >= 2));

plot(x, y);
grid on;
```