4.7절

4주1강. 확률변수와 확률분포3

복습

이산형 확률변수

- 이산형 확률변수란 0이 아닌 확률 값을 갖는 실수 값이 셀 수 있는 경우를 말한다. 즉, 확률변수가 이산점에서 0이 아닌 확률을 갖는 확률변수를 말한다.
- 이산형 확률변수 X의 가능한 값을 $x_1, x_2, \cdots, x_n$이라 하고 $X = x_i$일 때의 확률이 P_i라면 다음과 같이 표현한다.

$$P_r(X = x_i) = P_i, \qquad i = 1, 2, 3, \cdots, n$$

- **이산형 확률변수의 확률조건**

1. $0 \leq P_i \leq 1,\ \ i = 1, 2, 3, \cdots, n$. 즉, 각 x_i가 나타날 확률은 0과 1사이의 값을 갖는다.
2. $\sum_{i=1}^{n} P_i = 1$. 즉, 모든 가능한 경우의 확률의 합은 1이다.

확률 질량 함수

- 이산형 확률변수는 이산점(discrete points)에서 0이 아닌 확률 값을 가지며 확률은, $P_r(X = x_i) = P_i, \;\; i = 1, 2, 3, \cdots, n$으로 표현한다. 이와 같이 각 이산점에 있어서 확률의 크기를 표현하는 함수를 **확률질량함수(probability mass function)**라고 한다.

연속형 확률변수

- **연속형 확률변수**란?
  가능한 값이 실수의 어느 특정구간 전체에 해당하는 확률변수를 말한다.
  즉, 특정 실수 구간에서 0이 아닌 확률을 갖는 확률변수이다.

- 연속형 확률변수는 특정한 실수구간 내에서 0이 아닌 확률을 가지므로 이 구간에 대한 확률은 함수의 형태로 표현한다.

확률밀도함수

- 연속형 확률변수 $\boldsymbol{X}$ 의 확률함수를 $\boldsymbol{f(x)}$라고 할 때, $\boldsymbol{f(x)}$는 **확률밀도함수 (probability density function; p.d.f)**라고 부르며 다음 조건을 만족한다.

1. 모든 $\boldsymbol{X}$ 값에 대하여 $\boldsymbol{f(x)} \geq 0$이다. 즉, $\boldsymbol{X}$의 모든 실수 값에 대하여 확률밀도함수는 0 이상이다.
2. $\boldsymbol{X}$의 모든 가능한 값의 확률은 적분 $\int_{-\infty}^{\infty} f(x)\,dx$로 구하며 이 값은 항상 1이다.
3. 구간 $(\boldsymbol{a}, \boldsymbol{b})$의 확률은 $\int_{a}^{b} f(x)\,dx$이다. 즉 구간 $(\boldsymbol{a}, \boldsymbol{b})$에 대한 $\boldsymbol{X}$의 확률은 그 구간에 있어서 확률밀도함수 $\boldsymbol{f(x)}$로 만들어지는 면적의 크기이다.

누적분포함수

- **누적분포함수(cumulative distribution function ; c.d.f)**란?

특정값 a에 대하여 확률변수 X가 $X \le a$인 모든 경우의 확률의 합으로 다음과 같이 표현한다.

$$P_r(X \le a) = F_X(a)$$

- 이산형 확률변수에서는 $F_X(a) = \sum_{x_i \le a} P(X = x_i)$이고,

- 연속형 확률변수에서는 $F_X(a) = \int_{-\infty}^{a} f(x)\,dx$이다.

결합 확률 분포

1. 이산형 확률변수인 경우

- 확률변수 X의 가능한 값이 $x_1, \cdots, x_n$이고, 확률변수 Y의 가능한 값이 $y_1, \cdots, y_m$이라면, (X, Y)의 결합확률분포는 다음과 같이 정의한다.

$$P_r(X = x_i, Y = y_j) = P_{ij}, \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, m$$

- 여기에서 함수 $P_r(X = x_i, Y = y_j) = P_{ij}$를 **결합확률질량함수(joint probability mass function)**라고 정의하며, 모든 (i, j)에 대해 다음이 성립한다.

$$P_{ij} \geq 0, \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} = 1$$

결합 확률 분포

2. 연속형 확률변수인 경우

- 확률변수 X, Y의 **결합확률밀도함수(joint probability density function)**를 $f(x,y)$라고 정의할 때, 모든 (x,y)에 대해 다음이 성립한다.

$$f(x,y) \geq 0$$

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

주변확률분포의 정의

1. 이산형 확률변수

- 확률변수 X의 가능한 값이 $x_1, \cdots, x_n$이고, 확률변수 Y의 가능한 값이 $y_1, \cdots, y_m$이며, X와 Y의 결합확률분포가 $P_r(X = x_i, Y = y_j) = P_{ij},\ (i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m)$이라고 할 때, X와 Y의 주변확률분포는 각각 다음과 같이 구한다.
- X의 **주변확률분포** :

$$P(X = x_i) = \sum_{j=1}^{m} P_{ij} = P_{i1} + P_{i2} + \cdots + P_{im}$$

- Y의 **주변확률분포** :

$$P(Y = y_j) = \sum_{i=1}^{n} P_{ij} = P_{1j} + P_{2j} + \cdots + P_{nj}$$

결합 확률 분포

2. 연속형 확률변수인 경우

- 두 확률변수 X, Y의 결합확률밀도함수를 $f(x,y)$라고 할 때, X, Y의 주변확률분포는 각각 다음과 같이 정의한다.
- X의 **주변확률분포** :

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

- Y의 **주변확률분포** :

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

- 연속형 확률분포에서 주변확률분포함수를 **주변확률밀도함수(marginal probability density function)**라고 부른다

두 확률변수의 독립성

1. 이산형 확률변수

- 두 이산형 확률변수 X, Y의 결합확률분포함수를 $P(X = x, Y = y)$ 라 하고, 각각의 주변분포함수를 $P(X = x)$, $P(Y = y)$라고 할 때, 모든 가능한 (x, y) 값에 대하여,

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

일 때 한하여 두 변수 X 와 Y 는 **독립**이다.

2. 연속형 확률변수

- 두 연속형 확률변수 X, Y의 결합확률밀도함수를 $f(x, y)$라 하고, $f(x)$와 $f(y)$를 각각 X, Y의 주변확률밀도함수라고 하면,

$$f(x, y) = f(x) \cdot f(y)$$

일 때 한하여 두 변수 X 와 Y 는 **독립**이다.

4.7절

4.7. 평균, 분산과 공분산

기댓값(평균)

- **기대값**(expected value) 또는 **평균**(mean, average)이라는 개념은 확률분포에서 **분포의 무게중심**을 말하며, 확률 값을 가중치로 하는 확률변수의 가능한 값에 대한 가중평균(weighted average)이라고 할 수 있다.

1. **이산형 확률변수**

- 이산형 확률변수 X의 가능한 값이 $(x_1, x_2, \cdots, x_n)$이며, $P(X = x_i) = P_i,\ i = 1, 2, \cdots, n$일 때, X 의 기대값 $E(X)$는 다음과 같다.

$$E(X) = \sum_{i=1}^{n} x_i P_i$$

2. **연속형 확률변수**

- 연속형 확률변수 X의 확률밀도함수를 $f(x)$라 하면, X의 기대값 $E(X)$는 다음과 같다.

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

기댓값의 특성

- X, Y를 확률변수, a, b를 상수라고 할 때, 기댓값은 항상 다음 조건을 만족한다.

1. $E(a) = a$
2. $E(aX + b) = aE(X) + b$
3. $E(aX + bY) = aE(X) + bE(Y)$

- 상수 a와 b를 제외할 때 $E(X + Y) = E(X) + E(Y)$가 되는데, 이는 **두 확률변수의 합의 기대값은 각 확률변수의 기대값의 합과 같음을 의미**한다.

분산

- 평균이 확률분포의 무게중심인 데 비하여 분산은 **확률분포의 흩어진 정도를 측정**하는 것으로 평균이 같은 경우에도 분산의 크기에 따라서 분포의 모양이 달라진다.

- 확률변수 X의 **분산(variance)**은 $E(X) = \mu$라고 할 때, X와 μ의 편차의 제곱, 즉 $(X-\mu)^2$의 기대값으로 $Var(X)$ 또는 σ_X^2으로 표현되며,

$$\sigma_X^2 = E(X-\mu)^2$$

- 으로 구한다. 여기에서 확률변수 X를 나타낼 필요가 없는 경우에는 σ_X^2에서 X를 없애고 σ^2으로 표현한다. 분산의 양의 제곱근을 **표준편차(standard deviation)**라고 하며 σ로 표현한다. 즉, $\sigma = \sqrt{\sigma^2}$이다.

분산의 계산공식

- $\boldsymbol{\sigma^2 = E(X^2) - (E(X))^2 = E(X^2) - \mu^2}$

$$\begin{aligned} E(X-\mu)^2 &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu\cdot\mu + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

- $E(X^2)$의 계산

$$E(X^2) = \sum_{i=1}^{n} x_i^2 P_i \qquad (X : \text{이산형 확률변수})$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \qquad (X : \text{연속형 확률변수})$$

공분산

- 공분산은 두 확률변수의 결합분포를 알고 있는 경우에 구할 수 있는 모수이다. 주로 두 변수 사이의 관계의 밀접도를 측정하는 상관계수(correlation coefficient)를 구하는 과정에서 계산되는 경우가 많다.

- 두 확률변수 X와 Y의 **공분산(covariance)**은 $Cov(X,Y)$ 또는 σ_{XY}로 표현하며 다음과 같이 계산한다.

$$\boldsymbol{\sigma_{XY}} = \boldsymbol{E}[(\boldsymbol{X}-\boldsymbol{\mu_X})(\boldsymbol{Y}-\boldsymbol{\mu_Y})] = E(X\cdot Y - Y\mu_X - X\mu_Y + \mu_X\cdot\mu_Y)$$

$$= E(XY) - \mu_X\cdot\mu_Y - \mu_Y\cdot\mu_X + \mu_X\cdot\mu_Y$$

$$= \boldsymbol{E}(\boldsymbol{XY}) - \boldsymbol{\mu_X}\cdot\boldsymbol{\mu_Y}$$

두 확률변수의 독립성과 공분산

- 두 확률변수 X, Y가 서로 독립이면 두 확률변수의 공분산은 0이다. 즉, $\boldsymbol{\sigma_{XY} = 0}$이다. 그러나 두 확률변수의 공분산이 0이라고 할지라도 두 확률변수가 항상 독립인 것은 아니다.

상관 계수

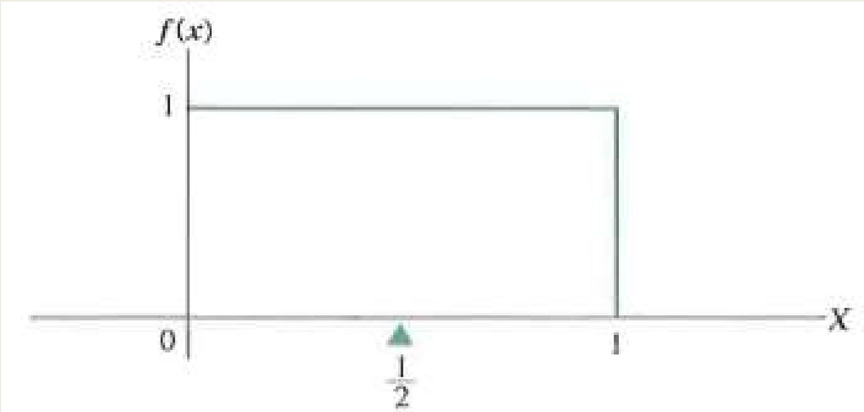
- 상관계수 ρ는 두 확률변수 사이의 관계의 밀접도를 측정하는 값으로 이용된다.
- $\rho = 0$: ρ의 계산에서 σ_{XY}=0임을 의미하며 이는 X와 Y가 서로 독립인 경우에 구할 수 있다.
- $\rho = 1$(또는 $\rho = -1$) : ρ의 계산에서 $\sigma_{XY} = \sigma_X \cdot \sigma_Y$임을 의미하며 X와 Y의 관계에서 $Y = aX + b$ (a, b는 상수, $a \neq 0$)와 같은 선형식으로 표현될 때 구하여진다.
- 두 확률변수 X와 Y에 대하여 σ_X^2, σ_Y^2을 각각의 분산이라 하고, σ_{XY}를 X와 Y의 공분산이라고 할 때, X와 Y의 **상관계수(correlation coefficient)**는 ρ로 표현하며, 다음과 같이 정의한다.

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

예4-17 연속형 확률변수 X의 분포가 다음과 같을 때 X의 확률분포를 그림으로 표현해보자.

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{다른 곳에서} \end{cases}$$

(sol)



이 그림에서 무게중심이 1/2 임을 알 수 있다.

예4-18

- 복권 100장을 발매한 후에 임의로 추첨하여 1등 1명에게 10,000원, 2등 3명에게 5,000원, 3등 10명에게 1,000원의 상금을 지급한다고 할 때 복권 한 장의 기대가치는 얼마인가?

(sol) 복권상금의 확률분포는 다음 표와 같다.

X	10,000	5,000	1,000	0	합
확률	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{10}{100}$	$\frac{86}{100}$	1

따라서 복권 1장의 기대가치 E(X)는 다음과 같이 계산된다.

$$E(X) = 10{,}000 \times \frac{1}{100} + 5{,}000 \times \frac{3}{100} + 1{,}000 \times \frac{10}{100} + 0 \times \frac{86}{100} = 350$$

예4-19. 두 확률변수 X와 Y의 분포가 다음과 같을 때, $E(X+Y)=E(X)+E(Y)$임을 보여라.

X \ Y	0	1	합
1	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{5}{18}$
2	0	$\frac{2}{9}$	$\frac{2}{9}$
3	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
합	$\frac{1}{3}$	$\frac{2}{3}$	1

(sol)

$E(X)=1\times\frac{5}{18}+2\times\frac{2}{9}+3\times\frac{1}{2}=\frac{40}{18}=\frac{20}{9}$ $E(Y)=0\times\frac{1}{3}+1\times\frac{2}{3}=\frac{2}{3}$ $E(X)+E(Y)=\frac{20}{9}+\frac{2}{3}=\frac{26}{9}$

확률변수 Z를 $Z=X+Y$로 정의하면 Z의 분포와 평균은 다음과 같다.

Z	1	2	3	4	합
확률	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{7}{18}$	$\frac{1}{3}$	1

$$E(Z)=1\times\frac{1}{6}+2\times\frac{1}{9}+3\times\frac{7}{18}+4\times\frac{1}{3}=\frac{52}{18}$$

따라서 $E(X+Y)=E(X)+E(Y)$임을 알 수 있다.

예4-20. 이산형 확률변수 X의 확률분포가 다음과 같을 때 X의 평균과 분산을 계산하라.

X	0	1	2	3	합
확률	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

(sol) $\mu = E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5$

$$E(X^2) = \sum x_i^2 P_i$$
$$= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 4 \times \frac{3}{8} + 9 \times \frac{1}{8}$$
$$= \frac{3+12+9}{8} = \frac{24}{8} = 3$$

$$\sigma^2 = E(X^2) - \mu^2 = 3 - 2.25 = 0.75$$

예4-21. 연속형 확률변수 X가 다음과 같은 확률밀도함수를 가질 때 X의 평균과 분산을 계산하라.

$$f(x) = \begin{cases} \frac{1}{2}x, & 0 \leq x \leq 2 \\ 0, & \text{다른 곳에서} \end{cases}$$

(sol) $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^2 x \cdot \frac{1}{2}x\,dx = \frac{1}{2}\left[\frac{1}{3}x^3\right]_0^2 = \frac{1}{2} \cdot \frac{8}{3} = \frac{4}{3}$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^2 x^2 \cdot \frac{1}{2}x\,dx = \frac{1}{2}\int_0^2 x^3\,dx$$
$$= \frac{1}{2}\left[\frac{1}{4}x^4\right]_0^2 = \frac{1}{2} \times \frac{1}{4} \times 16 = 2$$

$$\sigma^2 = E(X^2) - \mu^2 = 2 - \frac{16}{9} = \frac{2}{9}$$

예4-22 두 이산형 확률변수의 분포가 다음과 같을 때 X와 Y의 분산과 공분산을 구하라.

X \ Y	0	1	2	합
1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{3}$
3	$\frac{1}{12}$	$\frac{1}{2}$	$\frac{1}{12}$	$\frac{2}{3}$
합	$\frac{1}{4}$	$\frac{7}{12}$	$\frac{1}{6}$	1

(sol) X와 Y의 평균과 분산은 다음과 같이 계산된다.

$$\mu_X = E(X) = 1 \times \frac{1}{3} + 3 \times \frac{2}{3} = \frac{7}{3}$$

$$E(X^2) = 1 \times \frac{1}{3} + 9 \times \frac{2}{3} = \frac{19}{3}$$

$$\sigma_X^2 = E(X^2) - \mu_X^2 = \frac{19}{3} - \frac{49}{9} = \frac{8}{9}$$

$$\mu_Y = E(Y) = 0 \times \frac{1}{4} + 1 \times \frac{7}{12} + 2 \times \frac{1}{6} = \frac{11}{12}$$

$$E(Y^2) = 0 \times \frac{1}{4} + 1 \times \frac{7}{12} + 4 \times \frac{1}{6} = \frac{15}{12} = \frac{5}{4}$$

$$\sigma_Y^2 = E(Y^2) - \mu_Y^2 = \frac{5}{4} - \frac{121}{144} = \frac{59}{144}$$

$E(XY)$와 공분산은 다음과 같이 계산한다.

$$E(XY) = \sum\sum x_i \cdot y_j \cdot P_{ij} = 0 + \frac{1}{12} + \frac{1}{6} + 0 + \frac{3}{2} + \frac{1}{2} = \frac{9}{4}$$

$$\sigma_{XY} = E(XY) - \mu_X \cdot \mu_Y = \frac{9}{4} - \frac{7}{3} \times \frac{11}{12} = \frac{1}{9}$$

예4-23. 연속형 확률변수 X와 Y의 결합확률밀도함수가 다음과 같을 때 X, Y의 공분산을 구하라.

- $f(x,y) = \begin{cases} 1, & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & \text{다른 곳에서} \end{cases}$

(sol) X와 Y의 주변밀도함수는 각각 다음과 같다.

$$f(x) = \int f(x,y)\,dy = \int_0^1 1\,dy = 1, 0 \le x \le 1$$

$$f(y) = \int f(x,y)\,dx = \int_0^1 1\,dx = 1, 0 \le y \le 1$$

X와 Y의 평균과 공분산은 다음과 같이 계산된다.

$$\mu_X = E(X) = \int_0^1 xf(x)\,dx = \int_0^1 x\,dx = \frac{1}{2} \qquad \mu_Y = E(Y) = \int_0^1 yf(y)\,dy = \int_0^1 y\,dy = \frac{1}{2}$$

$$E(XY) = \iint x \cdot y \cdot f(x,y)\,dxdy = \int_0^1 \int_0^1 xy\,dxdy = \int_0^1 y \left[\int_0^1 x\,dx\right] dy = \int_0^1 \frac{1}{2} y\,dy = \frac{1}{4}$$

$$\sigma_{XY} = E(XY) - \mu_X \cdot \mu_Y = \frac{1}{4} - \frac{1}{2} \times \frac{1}{2} = 0$$

예4-24. [예 4.22]에 있는 X, Y의 분포를 이용하여 X와 Y의 상관계수를 구하라.

$X \backslash Y$	0	1	2	합
1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{3}$
3	$\frac{1}{12}$	$\frac{1}{2}$	$\frac{1}{12}$	$\frac{2}{3}$
합	$\frac{1}{4}$	$\frac{7}{12}$	$\frac{1}{6}$	1

(sol) X와 Y의 평균과 공분산은 다음과 같이 계산된다.

$$\sigma_X^2 = \frac{8}{9} \qquad \sigma_Y^2 = \frac{59}{144} \qquad \sigma_{XY} = \frac{1}{9}$$

상관계수 ρ는 다음과 같이 계산된다.

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}} = \frac{\frac{1}{9}}{\sqrt{\frac{8}{9} \times \frac{59}{144}}} = 0.184$$

끝~~ ♥♥