2강 확률분포(1)

학습내용
1. 확률분포함수를 이해한다.
2. 기댓값과 분산을 이해한다.
3. 마코프 부등식을 이해한다.
4. 적률생성함수를 이해한다.

01 확률분포함수

1. 확률변수

- 이산형 확률변수
  - 확률변수의 가능한 값이 유한 개이거나 셀 수 있는 경우
  - 확률질량함수: f(x) = P(X=x)
  - 확률질량함수의 성질
    - $0 \le f(x) \le 1$
    - $\sum_{x=1}^{n} f(x) = 1$
    - $P(a \le X \le b) = \sum_{a}^{b} f(x)$
- 예) 불량품 3개, 양품 2개인 상자에서 2개의 제품을 꺼낼 때 불량품 수의 확률질량함수는?
  - $f(x) = \frac{{}_3C_x \cdot {}_2C_{2-x}}{{}_5C_2}, \quad x = 0, 1, 2$
- 연속형 확률변수
  - 확률변수의 가능한 값이 무한 개이며 셀 수 없는 경우
  - 연속형 확률변수의 분포를 결정하는 함수: 누적분포함수
    - $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt, \quad f(x) = \frac{d}{dx}F(x)$
  - 누적분포함수를 미분하면 확률밀도함수가 된다.
- 예) 10분 간격으로 출발하는 버스를 기다리는 시간의 확률밀도함수는?
  - $F(x) = P(X \le x) = \frac{x-0}{10} = \frac{x}{10} = \int_{0}^{x} \frac{1}{10}dt$
  - 그런데,

    $f(x) = \frac{d}{dx}F(x)$

    이므로

$$f(x) = \begin{cases} \frac{1}{10} & \text{if} \quad 0 \le x \le 10 \\ 0 & \text{if} \quad x < 0 \;\; or \;\; x > 10 \end{cases}$$

이다.

- 확률밀도함수의 성질
  - $0 \le f(x)$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
  - $P(a \le X \le b) = \int_{a}^{b} f(x)dx$

- 예) 다음 확률밀도함수에서 C와 P(1<x<2)를 구하시오.
  - $$f(x) = \begin{cases} Ce^{-x} & \text{if} \;\; x \ge 0 \\ 0 & \text{if} \;\; x < 0 \end{cases}$$
    $$\int_{-\infty}^{\infty} f(x)dx = 1$$
  - 인데, x<0 인 경우 f(x)=0 이므로
    $$\int_{0}^{\infty} f(x)dx = \int_{0}^{\infty} Ce^{-x}dx = 1$$
    이다.
  - $\int_{0}^{\infty} Ce^{-x}dx = C\int_{0}^{\infty} e^{-x}dx = C\left[-e^{-x}\right]_{0}^{\infty} = C\left[0-(-1)\right] = C = 1$
  - C가 1이므로
  - $P(1 < x < 2) = \int_{1}^{2} e^{-x}dx = \left[-e^{-x}\right]_{1}^{2} = -e^{-2} - (-e)^{-1} = e^{-1} - e^{-2}$

2. 기댓값

- 확률분포
  - 확률변수 모든 값에 대응하는 확률을 나타내는 분포
    - 이산형 확률변수: 확률질량함수
    - 연속형 확률변수: 확률밀도함수
- 기댓값
  - 분포의 무게중심(균형점)
    - 이산형 확률변수:
      $$E(X) = \sum_{x=0}^{\infty} xf(x)$$
    - 연속형 확률변수:
      $$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$
- 예) 동전을 2번 던져서 나온 앞면의 수의 기댓값은?

X	0	1	2
E(X)	1/4	1/2	1/4

$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

- 예) 다음 확률밀도함수를 가지는 X의 기댓값은?

$$f(x) = \begin{cases} 2(1-x) & \text{if } \ 0 \le x \le 1 \\ 0 & \text{if } \ x < 0, x > 1 \end{cases}$$

$$E(X) = \int_0^1 x \cdot 2(1-x)dx = \int_0^1 (2x - 2x^2)dx$$
$$= \left[2 \cdot \frac{1}{2}x^2 - 2 \cdot \frac{1}{3}x^3\right]_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

3. 기댓값의 성질

- 기댓값의 성질
  - E(aX+b) = aE(X) + b
  - E(g(X)+h(X)) = E(g(X)) + E(h(X))

03 분산과 표준편차

1. 분산

- 분산(variance): 확률변수의 흩어져 있는 정도
  - $Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$
- 분산의 성질
  - $Var(aX + b) = a^2 Var(X)$

2. 표준편차

- 표준편차(standard deviation)
  - $Sd(X) = \sqrt{Var(X)}$
- 예) 동전을 2번 던져서 나온 앞면의 수의 분산과 표준편차는?

X	0	1	2
f(x)	1/4	1/2	1/4

$$E(X^2) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4} = \frac{3}{2}$$

$$Var(X) = E(X^2) - (E(X)^2) = \frac{3}{2} - 1 = \frac{1}{2}$$

$$Sd(X) = \sqrt{Var(X)} = \sqrt{\frac{1}{2}}$$

- 예) 다음 확률밀도함수를 가지는 X의 분산과 표준편차는?

  $$f(x) = \begin{cases} 2(1-x) & \text{if } \ 0 \le x \le 1 \\ 0 & \text{if } \ x < 0, x > 1 \end{cases}$$

$$E(X) = \int_0^1 x \cdot 2(1-x)dx = \int_0^1 (2x - 2x^2)dx$$

$$= \left[2 \cdot \frac{1}{2}x^2 - 2 \cdot \frac{1}{3}x^3\right]_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

$$Var(X) = E(X^2) - (E(X)^2) = \int_0^1 x^2 \cdot 2(1-x)dx - \frac{1}{3}^2$$

$$= \int_0^1 (2x^2 - 2x^3)dx - \frac{1}{9}$$

$$= \int_0^1 (2x^2 - 2x^3)dx - \frac{1}{9} = \left[2 \cdot \frac{1}{3}x^3 - 2 \cdot \frac{1}{4}x^4\right]_0^1 - \frac{1}{9}$$

$$= \frac{2}{3} - \frac{2}{4} - \frac{1}{9} = \frac{12 - 9 - 2}{18} = \frac{1}{18}$$

04 마코프 부등식

1. 체비세프 부등식

- 체비세프(Chebyshev) 부등식
  - 확률변수
    $X \sim (\mu, \sigma^2), \;\; \sigma^2 < \infty$
    , *k*: 양의 상수
    $\Rightarrow P(|X-\mu| \geq k) \leq \dfrac{\sigma^2}{k^2}$
- 증명

$$P(|X-\mu| \geq k) = \int_{|X-\mu|\geq k} f(x)dx = \int I_{|X-\mu|\geq k} f(x)dx$$

$$= E(I_{|X-\mu|\geq k}) \leq E(I_{|X-\mu|\geq k} \cdot \frac{(X-\mu)^2}{k^2}) \leq E(\frac{(X-\mu)^2}{k^2}) = \frac{\sigma^2}{k^2}$$

2. 마코프 부등식

- 마코프(Markov) 부등식
  - 체비셰프 부등식을 일반화 한 것
  - 확률변수
    $X, \;\; E(|X|^r) < \infty, \;\; r, k > 0$

    $\Rightarrow P(|X| \geq k) \leq \dfrac{E(|X|^r)}{k^r}$
- r=1 일 때 증명
  $P(|X| \geq k) \leq \dfrac{E(X)}{k}$

$$E(X) = \int_0^\infty xf(x) = \int_0^k xf(x)dx + \int_k^\infty xf(x)dx$$

$$\geq \int_0^k xf(x)dx + k\int_0^k f(x)dx \geq k\int_0^k f(x)dx = kP(X \geq k)$$

3. 분산의 계산

- 예) X가 다음 확률밀도함수를 가질 때
  $P(|X-\mu_X|) \geq \dfrac{3}{2}\sigma_X$
  의 상한값은?

$$f(x) = \frac{1}{2\sqrt{3}}, \;\; -\sqrt{3} < x < \sqrt{3}$$

- 체비셰프 부등식을 활용하려면 σ_X^2 을 계산해야 한다.(그래야 *k*를 알 수 있으니까)
  - σ_X^2 을 계산하려면 μ_X 를 계산해야 한다.

$$\mu_X = \int_{-\sqrt{3}}^{\sqrt{3}} x \cdot \frac{1}{2\sqrt{3}} dx = \left[\frac{1}{2}x^2 \cdot \frac{1}{2\sqrt{3}}\right]_{-\sqrt{3}}^{\sqrt{3}} = 0$$

- μ_X =0 이므로,

$$\sigma_X^2 = Var(X) = E(X^2) - (E(X)^2) = E(X^2)$$
$$= \int_{-sqrt3}^{\sqrt{3}} x^2 \cdot \frac{1}{2\sqrt{3}} dx = \left[\frac{1}{3}x^3 \cdot \frac{1}{2\sqrt{3}}\right]_{-\sqrt{3}}^{\sqrt{3}}$$
$$= \left[\frac{3\sqrt{3}}{3} - (-\frac{3\sqrt{3}}{3}) \cdot \frac{1}{2\sqrt{3}}\right] = 1$$

- 주어진 식 $P(|X - \mu_x|) \geq \frac{3}{2}\sigma_X$ 에서 *μ*=0, *σ*=1 이라고 하면,
  - $k = \frac{3}{2}$ 이다.
- 체비셰프 부등식: $P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$ 과 $\sigma_X = 1, k = \frac{3}{2}$ 이라는 사실을 적용하면,
  - $P(|X - \mu_X| \geq \frac{3}{2}) \leq \frac{1^2}{(\frac{3}{2})^2} = \frac{4}{9}$ 임을 알 수 있다.
- 따라서, 상한값은 4/9 이다.

05 적률생성함수

1. 적률생성함수의 정의

- k차 적률(moment)
    - $$E(X^k) = \begin{cases} \sum_x x^k f(x), & discrete \\ \int_{-\infty}^{\infty} x^k f(x)dx, & continuous \end{cases}$$
    - 1차 적률: 기대값
    - 2차 적률: 분산(흩어짐)
    - 3차 적률: 왜도(치우침)
    - 4차 적률: 첨도(뾰족한 정도)
- 적률생성함수(moment generating function, mgf)
    - 0을 포함한 열린 구간의 t에 대해서
      $E(e^{tk}) < \infty$
      일 때,
      $$M(t) = E(e^{tk}) = \begin{cases} \sum_x e^{tk} f(x), & discrete \\ \int_{-\infty}^{\infty} e^{tk} f(x)dx, & continuous \end{cases}$$

2. 적률생성함수의 성질

- 적률생성함수가 존재하면 모든 적률이 존재
    - $M(0) = 1$
    - $M'(0) = E(X)$
    - $M^{(k)}(0) = E(X^k)$
- 두 확률변수 X, Y의 적률생성함수가 존재하고, 0을 포함한 열린 구간에서 일치하며 두 확률변수 X, Y의 분포가 일치
    - $M_x(t) = M_y(t), \ \ \forall t : -h < t < h(h > 0)$
  $\Leftrightarrow f_x(x) = f_y(y)$

3. 적률생성함수의 예

- 예) 확률변수 X의 확률분포가 다음과 같을 때 X의 적률생성함수와 기댓값은?

	0	**1**	**2**	**3**
f(x)	1/6	1/3	1/3	1/6

- 풀이)
    - 적률생성함수는

$$M(t) = E(e^{tx}) = e^{t\cdot 0}\cdot\frac{1}{6} + e^{t}\cdot\frac{1}{3} + e^{2t}\cdot\frac{1}{3} + e^{3t}\cdot\frac{1}{6}$$

$$= \frac{1}{6} + \frac{1}{3}e^{t} + \frac{1}{3}e^{2t} + \frac{1}{6}e^{3t}$$

- 적률생성함수의 1차미분식에 t=0을 대입하여 기대값을 구한다

  $$M'(0) = \left[\frac{1}{3}e^{t} + \frac{1}{3}2e^{2t} + \frac{1}{6}3e^{3t}\right]_{t=0}$$

  $$= \frac{1}{3} + \frac{2}{3} + \frac{3}{6} = \frac{3}{2}$$

- 이것은

  $$E(X) = \sum_{x=0}^{3} xf(x)$$

  식을 활용하여 구한 기대값과 동일하다.

  $$E(X) = 0\cdot\frac{1}{6} + 1\cdot\frac{1}{3} + 2\cdot\frac{1}{3} + 3\cdot\frac{1}{6} = \frac{1}{3} + \frac{2}{3} + \frac{1}{2} = \frac{3}{2}$$

* 예) 확률변수 X의 확률분포가 다음과 같을 때 X의 적률생성함수와 기대값은?

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

* 풀이)
  - 적률생성함수는 다음과 같이 구한다.

    $$M(t) = E(e^{tx}) = \int_0^{\infty} e^{tx}\cdot e^{-x}dx$$

    $$= \int_0^{\infty} e^{x(t-1)}dx = \left[\frac{1}{t-1}e^{x(t-1)}\right]_{t=0}^{t=\infty}$$

    $$= 0 - \frac{1}{t-1} = \frac{1}{1-t}$$
  - 참고로 x에 곱해지는 t-1이 0보다 작아야만 x가 무한대일 때 전체 식의 값이 0이 되기 때문에 t-1<0 을 충족해야 한다.
  - 기대값을 구해보자

$$E(X) = [M'(t)]_{t=0} = \left[-\frac{-1}{(1-t)^2}\right]_{t=0} = 1$$

06 정리하기

- 확률분포는 이산형 확률변수의 경우 확률질량함수로, 연속형 확률변수의 경우 확률밀도함수로 결정된다.
- 기댓값은 확률분포의 무게중심(균형점)이다.
- 분산으로 확률분포가 흩어져 있는 정도를 측정한다.
- 마코프 부등식은 확률분포의 상한을 제시한다.
- 적률생성함수가 존재하면 확률변수의 적률을 구할 수 있고, 확률분포와 대응한다.