2장. 확률변수

확률변수와 분포함수

확률변수의 기댓값(=평균)

적률생성함수

체비셰프 부등식

확률변수(random variable)

$(\Omega, \mathcal{F}, P)$를 확률공간이라 하자.

(1) 함수 $X : \Omega \longrightarrow \mathbb{R}$를 Ω 위에서 정의되는 확률변수라 한다.

(2) 확률변수 $X : \Omega \longrightarrow \mathbb{R}$의 치공간:

$$R_X = \{X(\omega) \,|\, \omega \in \Omega\} \subset \mathbb{R}$$

(3) 임의의 $x \in \mathbb{R}$과 임의의 $B \in \mathfrak{B}$에 대하여,

$$[X = x] = \{\omega \in \Omega \,|\, X(\omega) = x\} = X^{-1}\Big(\{x\}\Big)$$
$$[X \leqslant x] = \{\omega \in \Omega \,|\, X(\omega) \leqslant x\} = X^{-1}\Big((-\infty, x]\Big)$$
$$[X \in B] = \{\omega \in \Omega \,|\, X(\omega) \in B\} = X^{-1}(B)$$

Example

한 개의 동전을 세 번 던지는 실험을 생각하자.

- 표본공간 Ω
- $\mathcal{F} = \mathcal{P}(\Omega)$
- P : 사전확률

확률변수 $X : \Omega \longrightarrow \mathbb{R}$, $X(\omega) := \omega$에서 앞면의 개수

(1) X의 치공간 R_X

(2) $P[X = 1]$

(3) $P[X \leqslant 1]$

(4) $P\Big[X \in (2, 5]\Big]$

(5) $P[X \leqslant -1]$

확률변수의 (누적)분포함수(distribution function)

X가 확률공간 $(\Omega, \mathcal{F}, P)$ 위에서 정의되는 확률변수일 때, 함수

$$F_X : \mathbb{R} \longrightarrow \mathbb{R}, \quad x \mapsto F_X(x) = P[X \leqslant x],$$

를 확률변수 X의 (누적)분포함수라 한다.

Theorem(분포함수의 기본 성질)

F_X가 X의 분포함수이면,

(1) F_X는 단조증가함수이다.

(2) $\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$

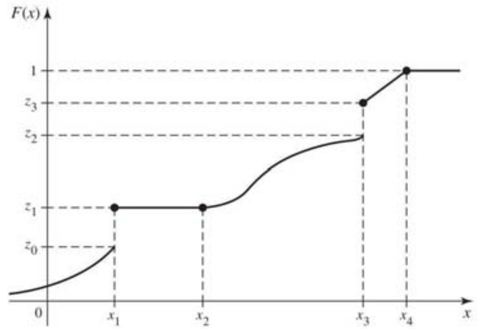
(3) $\lim_{\epsilon \to 0+} F_X(x + \epsilon) = F_X(x)$:우방 연속성(right continuous)

분포함수 $F_X(x)$의 역할

(1) $P[a < X \leqslant b] = P[X \leqslant b] - P[X \leqslant a] = F_X(b) - F_X(a)$

(2) $P[X = b] = \lim\limits_{h \to 0+} P[b - h < X \leqslant b] = F_X(b) - F_X(b - 0)$

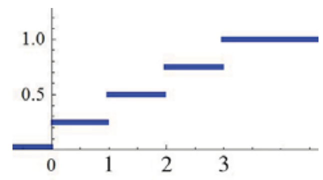
(3) $P[X < b] = F_X(b - 0)$



Example

확률변수 X의 분포함수가 다음과 같을 때, 다음의 확률을 구하시오.

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}, & 0 \leqslant x < 1 \\ \frac{1}{2}, & 1 \leqslant x < 2 \\ \frac{3}{4}, & 2 \leqslant x < 3 \\ 1, & x \geqslant 3 \end{cases}$$



(1) $P[0 < X \leqslant 2]$ (2) $P[0 \leqslant X \leqslant 2]$

(3) $P[0 < X < 2]$ (4) $P[0 \leqslant X < 2]$

(5) $P[1.1 < X < 2.9]$

이산확률변수, 연속확률변수

X가 확률공간 $(\Omega, \mathcal{F}, P)$ 위에서 정의되는 확률변수라 하자.

(1) X의 치공간 R_X가 가산일 때, X를 이산확률변수(discrete random variable)라 한다.

(2) 모든 $x \in \mathbb{R}$에 대하여 $P[X = x] = 0$일 때, X를 연속확률변수(continuous random variable)라 한다.

확률질량함수(probability mass function=pmf)

X가 확률공간 $(\Omega, \mathcal{F}, P)$ 위에서 정의되는 이산확률변수라 하자.

(1) 다음과 같이 정의되는 함수 f_X를 X의 확률질량함수라 한다.

$$f_X : \mathbb{R} \longrightarrow \mathbb{R}, \quad f_X(x) = P[X = x]$$

(2) 집합 $\{(x, f_X(x)) \mid x \in \mathbb{R}\}$을 X의 확률분포(probability distribution)라 한다.

Theorem(확률질량함수의 성질)

f_X가 이산확률변수 X의 확률질량함수일 때,

(1) $\forall x \in \mathbb{R}, \quad f_X(x) \geqslant 0$

(2) $\sum\limits_{x \in \mathbb{R}} f_X(x) = \sum\limits_{i \in I} f_X(x_i) = 1 \quad (R_X = \{x_i \mid i \in I\})$

Example

한 개의 동전을 앞면(H)이 나올 때까지 던지고 그때까지 던진 횟수를 X라 하자.(기하분포)

(1) 확률변수 X의 치공간
(2) 확률변수 X의 확률질량함수
(3) 확률변수 X의 분포함수
(4) 확률 $P[5 \leqslant X \leqslant 10]$

확률밀도함수(probability density function=pdf)

X가 확률공간 $(\Omega, \mathcal{F}, P)$ 위에서 정의되는 연속확률변수라 하자. 다음을 만족하는 음이 아닌 함수 $f_X : \mathbb{R} \longrightarrow \mathbb{R}$를 X의 확률밀도함수라 한다.

$$F_X(x) = P[X \leqslant x] = \int_{-\infty}^{x} f_X(t)\,dt$$

Theorem(확률밀도함수의 성질)

f_X가 연속확률변수 X의 확률밀도함수이면, 다음이 성립한다.

$$(1) \quad \forall a, b \in \mathbb{R}\ (a < b), \quad P[a < X \leqslant b] = \int_{a}^{b} f_X(t)\,dt$$

$$(2) \quad \int_{-\infty}^{\infty} f_X(t)\,dt = P[-\infty < X < \infty] = P(\Omega) = 1$$

Example

연속확률변수 X의 확률밀도함수 f_X가

$$f_X(x) = \begin{cases} ax^2, & 0 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

로 주어질 때, 상수 a와 확률 $P[\frac{1}{2} \leqslant X \leqslant 1]$을 구하시오.

Theorem(F_X와 f_X의 관계)

X가 확률변수이면 F_X는 f_X로부터 얻어지고, 그 역도 성립한다.

- X가 이산확률변수일 경우

$$F_X(x) = P[X \leqslant x] = \sum_{t \leqslant x} f_X(t)$$

$$f_X(x) = P[X = x] = F_X(x) - F_X(x - 0)$$

- X가 연속확률변수일 경우

$$F_X(x) = P[X \leqslant x] = \int_{-\infty}^{x} f_X(t)\,dt$$

$F_X(x)$가 미분가능한 x에 대하여

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{-\infty}^{x} f_X(t)\,dt$$

Example

연속확률변수 X의 분포함수 F_X가

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leqslant x \leqslant 1 \\ 1, & x > 1 \end{cases}$$

로 주어질 때, X의 확률밀도함수 f_X와 확률 $P[-2 < X \leqslant \frac{1}{2}]$을 구하시오.

Example

연속확률변수 X의 확률밀도함수 f_X가

$$f_X(x) = \begin{cases} \frac{x}{2}, & 0 < x < 2 \\ 0, & \text{otherwise} \end{cases} = \frac{x}{2} \cdot I_{(0,2)}(x)$$

로 주어질 때, X의 분포함수 F_X와 확률 $P[\frac{1}{2} < X \leqslant 1]$을 구하시오.

혼합형분포함수(Mixed Distribution Function)

X_d가 이산확률변수이고, X_c가 연속확률변수이고 각각의 분포함수를 F_{X_d}, F_{X_c}라 하면, 임의의 $0 < \alpha < 1$에 대하여

$$F_X(x) = \alpha F_{X_d}(x) + (1-\alpha) F_{X_c}(x)$$

를 분포함수로 하는 확률변수 X를 혼합확률변수라 한다.
여기서 α는 이산형 부분의 확률의 합이고, $1-\alpha$는 연속형 부분의 확률의 합이다.

Example(혼합형분포함수)

어느 건널목에 있는 신호등은 보행자를 기준으로 초록색이 1분, 빨간색이 3분 동안 켜진다고 한다. 어떤 사람이 이 건널목에 도착하여 기다리는 시간을 X라 할 때, X의 분포함수를 구하시오.

Solution. X의 치공간은 $R_X = \{0\} \cup (0, 3]$이다.

X_d = 초록색 신호등일 때 도착하면 기다리는 시간 (이산확률변수)
X_d의 분포함수:

$$F_{X_d}(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geqslant 0 \end{cases}$$

X_c = 빨간색 신호등일 때 도착하면 기다리는 시간 (연속확률변수)
X_c의 분포함수:

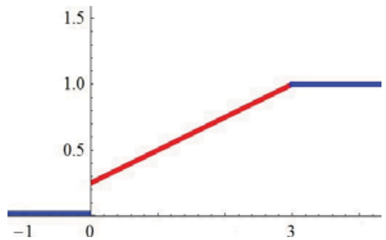
$$F_{X_c}(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{3}x, & 0 \leqslant x < 3 \\ 1, & x \geqslant 3 \end{cases}$$

X의 분포함수:

$$F_X(x) = \alpha F_{X_d}(x) + (1-\alpha) F_{X_c}(x) = \frac{1}{4} F_{X_d}(x) + \frac{3}{4} F_{X_c}(x)$$

$$= \frac{1}{4} \cdot \begin{cases} 0, & x < 0 \\ 1, & x \geqslant 0 \end{cases} + \frac{3}{4} \cdot \begin{cases} 0, & x < 0 \\ \frac{1}{3}x, & 0 \leqslant x < 3 \\ 1, & x \geqslant 3 \end{cases}$$

$$= \begin{cases} 0, & x < 0 \\ \frac{1}{4}x + \frac{1}{4}, & 0 \leqslant x < 3 \\ 1, & x \geqslant 3 \end{cases}$$



기댓값과 평균(Expectation and Mean)

f_X를 확률변수 X의 확률밀도(질량)함수라 하자.

$$E[X] = \begin{cases} \sum_{x\in\mathbb{R}} x \cdot f_X(x), & X\text{가 이산} \\ \int_{-\infty}^{\infty} x \cdot f_X(x)dx, & X\text{가 연속} \end{cases}$$

이 존재할 때 $E[X]$를 X의 기댓값 또는 평균이라 하고 μ_X로 나타낸다.

Remark
확률변수 X의 기댓값 $E[X]$는
$\sum_{x\in\mathbb{R}} |x| f_X(x) < \infty$, $\int_{-\infty}^{\infty} |x|\, f_X(x)dx < \infty$인 경우에만 정의한다.

Example

흰 공이 5개 검은 공이 4개가 있는 주머니에서 동시에 임의로 3개의 공을 꺼내서 그 속에 포함된 흰 공의 개수를 X라 하자.
(1) 확률변수 X의 확률질량함수를 구하시오.
(2) 확률변수 X의 평균을 구하시오.

Example

어떤 사람이 기다리는 e-메일이 오후 3시 이후에나 도착할 수 있다고 한다. 3시 이후 e-메일이 도착할 때까지의 시간을 X(단위는 시간)로 표시하면 그 확률밀도함수는 다음과 같이 가정된다.

$$f_X(x) = \begin{cases} \frac{1}{2}e^{-x/2}, & x \in (0,\infty) \\ 0, & \text{otherwise} \end{cases}$$

(1) e-메일이 3시 이후 30분 이내에 도착할 확률을 구하시오.
(2) 확률변수 X의 평균을 구하시오.

Theorem(확률변수의 함수의 기댓값)

f_X가 X의 확률밀도(질량)함수이고, $Y = h(X)$가 확률변수이면,

$$E[Y] = E\Big[h(X)\Big] = \begin{cases} \sum_{x\in\mathbb{R}} h(x) f_X(x), & X\text{가 이산} \\ \int_{-\infty}^{\infty} h(x) f_X(x)\, dx, & X\text{가 연속} \end{cases}$$

Proof. X를 이산확률변수, X의 치공간 $R_X = \{x_j \mid j \in J\}$라 하면
$Y = h(X)$는 이산확률변수, Y의 치공간 $R_Y = \{y_i \mid i \in I\}$

$$E[Y] = \sum_{y\in\mathbb{R}} y \cdot f_Y(y) = \sum_{y\in\mathbb{R}} y \cdot P[Y=y] = \sum_{i\in I} y_i \cdot P[Y=y_i]$$

한편, $[Y = y_i] = [h(X) = y_i]$이므로, $J_i := \{j \in J \mid h(x_j) = y_i\}$

$$J = \biguplus_{i\in I} J_i, \quad [Y = y_i] = [h(X) = y_i] = \biguplus_{j\in J_i} [X = x_j]$$

따라서

$$\begin{aligned} E[Y] &= \sum_{i\in I} y_i \cdot P[Y = y_i] = \sum_{i\in I} y_i \cdot P\left[\biguplus_{j\in J_i} [X = x_j]\right] \\ &= \sum_{i\in I} y_i \cdot \sum_{j\in J_i} P[X = x_j] \\ &= \sum_{i\in I}\sum_{j\in J_i} y_i \cdot P[X = x_j] \\ &= \sum_{i\in I}\sum_{j\in J_i} h(x_j) \cdot P[X = x_j] \\ &= \sum_{j\in J} h(x_j) \cdot P[X = x_j] \\ &= \sum_{x\in\mathbb{R}} h(x) \cdot P[X = x] \end{aligned}$$

Example

연속 확률변수 X의 확률밀도함수가

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

이라 하고, $Y = X^2$이라 하자.

(1) 확률변수 Y의 확률밀도함수를 구하시오.
(2) (1)의 결과를 이용하여 $E[Y]$를 구하시오.
(3) Theorem를 이용하여 $E[Y]$를 구하시오.

Theorem

확률변수 X와 임의의 함수 $Y = g(X)$, $Z = h(X)$에 대하여,
$E\Big[|X|\Big] < \infty$, $E\Big[|g(X)|\Big] < \infty$, $E\Big[|h(X)|\Big] < \infty$ 일 때,

(1) $E[X+Y] = E[X] + E[Y]$

(2) 임의의 $a, b \in \mathbb{R}$에 대하여, $E[aX+b] = aE[X] + b$

(3) $E\Big[g(X) + h(X)\Big] = E\Big[g(X)\Big] + E\Big[h(X)\Big]$

(4) $X \geqslant Y \geqslant 0$이면 $E[X] \geqslant E[Y]$

분산(variance)과 표준편차(standard deviation)

확률변수 X의 확률밀도(질량)함수를 f_X, 평균을 μ_X라 하자.

(1) $E\Big[(X-\mu_X)^2\Big]$이 존재할 때, 이 값을 X의 분산이라 하고 $\mathrm{Var}[X] = \sigma_X^2$로 나타낸다. 즉,

$$\mathrm{Var}[X] = E\Big[(X-\mu_X)^2\Big]$$

(2) $\mathrm{sd}[X] = \sigma_X = \sqrt{\mathrm{Var}[X]}$를 X의 표준편차라고 한다.

Theorem(분산의 계산식)

확률변수 X에 대하여, $E[X]$와 $E[X^2]$이 모두 존재하면,

$$\mathrm{Var}[X] = E[X^2] - E[X]^2$$

Theorem

f_X가 X의 확률밀도(질량)함수이고, $Y = h(X)$가 확률변수이면,
$E[Y] = \mu_Y$일 때, (단, 값이 존재할 때)

$$\mathrm{Var}[Y] = \mathrm{Var}\Big[h(X)\Big] = \begin{cases} \displaystyle\sum_{x\in\mathbb{R}} (h(x)-\mu_Y)^2 f_X(x), & X : \text{이산} \\ \displaystyle\int_{-\infty}^{\infty} (h(x)-\mu_Y)^2 f_X(x)\,dx, & X : \text{연속} \end{cases}$$

적률생성함수(moment generating function=mgf)

f_X를 확률변수 X의 확률밀도(질량)함수라 할 때,

(1) X의 r차 적률(r-th moment) $E[X^r]$: (단, 존재할 때)

$$E[X^r] = \begin{cases} \sum_{x\in\mathbb{R}} x^r f_X(x), & X\text{가 이산} \\ \int_{-\infty}^{\infty} x^r f_X(x)\,dx, & X\text{가 연속} \end{cases}$$

(2) X의 적률생성함수 $m_X(t)$: (단, 존재할 때)

$$m_X(t) = E\left[e^{tX}\right] = \begin{cases} \sum_{x\in\mathbb{R}} e^{tx} f_X(x), & X\text{가 이산} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx, & X\text{가 연속} \end{cases}$$

적률생성함수의 기능

(1) X의 적률생성함수 $m_X(t)$:

$$\begin{aligned} m_X(t) = E[e^{tX}] &= E\left[\sum_{n=0}^{\infty} \tfrac{1}{n!}(tX)^n\right] \\ &= E\left[1 + tX + \cdots + \tfrac{(tX)^n}{n!} + \cdots\right] \\ &\overset{(*)}{=} \sum_{n=0}^{\infty} E\left[\tfrac{1}{n!}(tX)^n\right] = E[1] + E[tX] + \cdots + E\left[\tfrac{(tX)^n}{n!}\right] + \cdots \\ &= 1 + E[X]\,t + \frac{E[X^2]}{2!}t^2 + \cdots + \frac{E[X^n]}{n!}t^n + \cdots \end{aligned}$$

(*) 등호가 성립되는 t. 예: $\sum_{n=0}^{\infty} E\left[\frac{1}{n!}|tX|^n\right]$가 수렴

(2) 적률생성함수 $m_X(t)$의 도함수: $m_X^{(n)}(0) = E[X^n]$

$$m_X'(t) = E[X] + E\left[X^2\right]t + \cdots + \tfrac{E[X^n]}{(n-1)!}t^{n-1} + \cdots$$

$$m_X''(t) = E\left[X^2\right] + E\left[X^3\right]t + \cdots + \tfrac{E[X^n]}{(n-2)!}t^{n-2} + \cdots$$

(3) 적률생성함수를 이용한 평균, 분산 구하기:

$$E[X] = m_X'(0)$$

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = m_X''(0) - \left(m_X'(0)\right)^2$$

Theorem(적률생성함수의 유일성)

$m_X(t)$와 $m_Y(t)$가 각각 확률변수 X, Y의 적률생성함수이고 적당한 $\delta > 0$에 대하여

$$\forall t \in (-\delta, \delta), \quad m_X(t) = m_Y(t)$$

이면, 모든 $x \in \mathbb{R}$에 대하여 $F_X(x) = F_Y(x)$이다.

Example

확률변수 X의 적률생성함수를 구하고 이를 이용하여 X의 평균과 분산을 구하시오.

(1) X의 확률질량함수

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \cdots, n \\ 0, & \text{otherwise} \end{cases}$$

(2) X의 확률밀도함수(단, $\beta > 0$)

$$f_X(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x \geqslant 0 \\ 0, & \text{otherwise} \end{cases}$$

마르코프 부등식(Markov Inequality)

(1) X가 음이아닌 확률변수이고 $E[X] < \infty$이면,

$$\forall a > 0, \quad P[X \geqslant a] \leqslant \frac{E[X]}{a}$$

(2) X가 확률변수이고, $p > 0$에 대하여 $E\Big[|X|^p\Big] < \infty$이면,

$$\forall a > 0, \quad P\Big[|X| \geqslant a\Big] \leqslant \frac{E\Big[|X|^p\Big]}{a^p}$$

체비셰프 부등식(Chebychev Inequality)

X가 확률변수이고, $E[X] = \mu,\ \mathrm{Var}[X] = \sigma^2$이면, 임의의 $k > 0$에 대하여

$$P\Big[|X - \mu| \geqslant k\sigma\Big] \leqslant \frac{1}{k^2}, \quad \text{즉,}$$

$$P\Big[|X - \mu| < k\sigma\Big] \geqslant 1 - \frac{1}{k^2}$$

Example

확률밀도(질량)함수를 알지 못하는 음이아닌 확률변수 X의 평균이 $E[X] = 50$이라 하자.

(1) 확률 $P[X \geqslant 75]$의 상계나 하계를 구하시오.

(2) X의 분산이 $\mathrm{Var}[X] = 25$일 때, 확률 $P[40 < X < 60]$의 상계나 하계를 구하시오.