통계 분석

2019년 2학기

강봉주

확률 분포

확률 분포

[확률 시행과 표본 공간]

실험(experiment) 또는 시행(trial):
가능한 모든 결과(outcome)가 정의되어 있고 무한히 반복 가능한 절차(procedure)를 의미한다. 무작위 또는 임의(random) 시행은 가능한 결과가 2개 이상인 시행이다.

1) 동전 던지기: 2개의 가능한 결과(베르누이 시행)
2) 주사위 던지기: 6개의 가능한 결과

확률 분포

[확률 시행과 표본 공간]

표본 공간(sample space):
확률 시행에서 가능한 모든 결과의 집합

1) 동전 던지기: {H, T}
2) 주사위 던지기: {1, 2, 3, 4, 5, 6}

확률 분포

[**사건과 상대도수**]

사건(event):
표본 공간의 부분 집합

1) 동전 던지기: $\{\phi, \{H\}, \{T\}, \{H, T\}\}$
2) 주사위 던지기: $\{\phi, \{1\}, \{2\}, \dots, \{1, \dots, 6\}\}$

확률 분포

[**사건과 상대도수**]

상대 도수(relative frequency):
N번 시행에서 특정 사건이 f번 일어났다고 한다면 상대 도수는 f/N

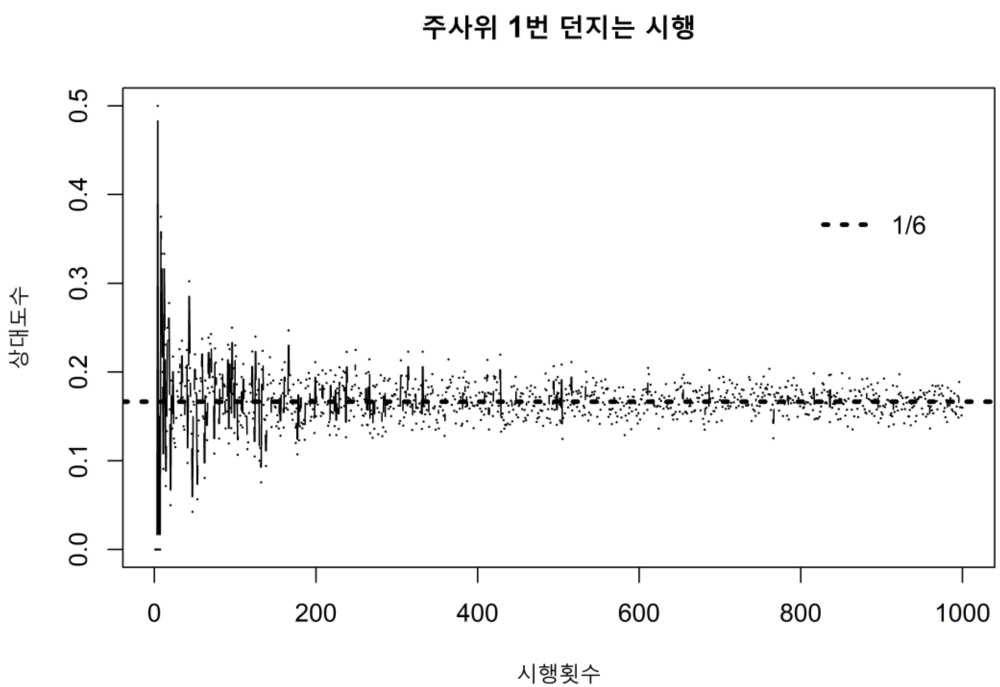
```
result <- c()
for (i in 1:1000){
  result[i] <- mean(rbinom(n=i, size=1, prob=1/6))
}
```

특정 사건의 1번 시행의 성공 확률이 1/6인 경우에 i개의 표본 추출

확률 분포

[사건과 상대도수]

상대 도수(relative frequency):
N번 시행에서 특정 사건이 f번 일어났다고 한다면 상대 도수는 f/N

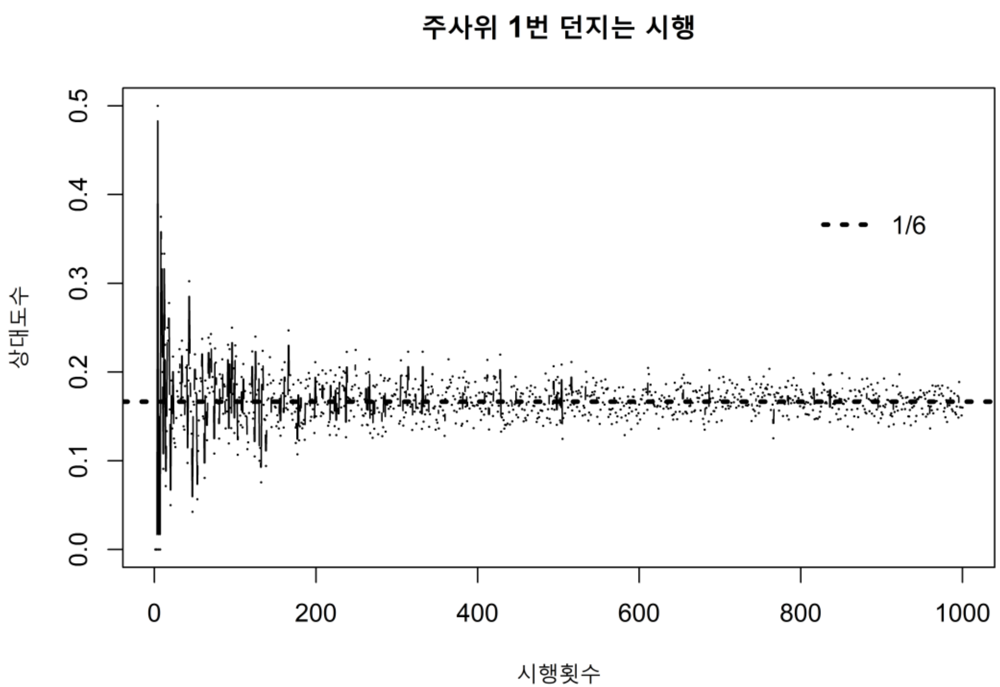


아주 많은 시행을 하면 특정 값으로 수렴

확률 분포

[**사건과 상대도수**]

예제)
다음 그림을 그려보자.



확률 분포

[**사건과 상대도수**]

많은 횟수를 시행하는 경우에는 상대 도수 값이 특정 값으로 안정화되는 경향이 있다. 그 값을 p 이라고 한다면 미래의 시행에서 해당 사건은 그 값만큼 일어날 것이라고 생각할 수 있다. 이 값을 사건 ω에 대한 확률(probability) 또는 확률 측도(probability measure)라고 한다. 이러한 방식으로 확률을 정의하는 것을 상대도수 접근 방법이라고 한다.

확률 분포

[사건과 상대도수]

$$\Pr(\omega) = \Pr(\{2\}) = P(\{2\})$$
$$\omega = \{2\} \in \mathcal{F} = \left\{\phi, \{1\}, \dots, \{1, \dots, 6\}\right\}$$

특정한 사건의 확률을 구한다는 것은 적절한 집합 함수 P를 찾는 것

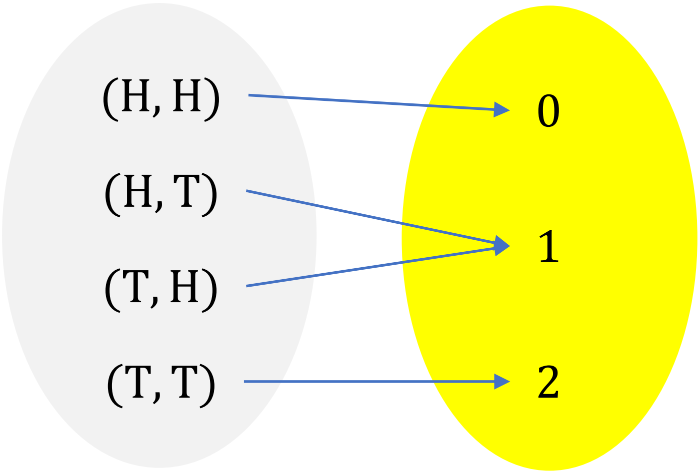
확률 분포

[확률 변수]

표본 공간(Ω)이 주어져 있을 때, 하나의 함수 X가 모든 $c \in \Omega$에 대하여 딱 한 개의 숫자만을 할당하는 경우에 즉, $X(c) = x$, 이 함수를 확률 변수라고 한다.

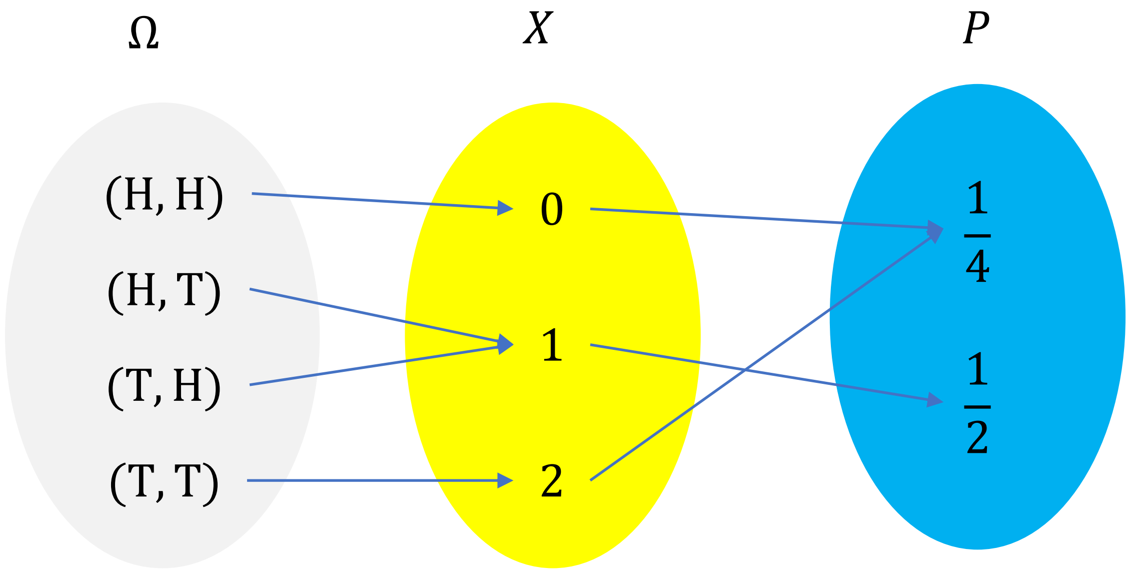
$$\Omega = \{(\mathrm{H},\mathrm{H}),(\mathrm{H},\mathrm{T}),(\mathrm{T},\mathrm{H}),(\mathrm{T},\mathrm{T})\}$$

X: 앞면이 나오는 개수



확률 분포

[확률 변수]



$$\Pr(\{(\mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{H})\}) = \Pr(X = 1) = \frac{1}{2}$$

표본 공간 위에서 확률을 정의하지 않고 숫자 값을 갖는 확률 변수 위에서 확률 정의

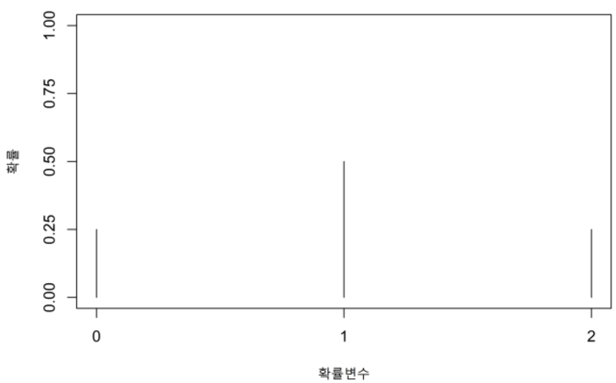
확률 분포

[확률 변수]

확률 분포표:

확률변수의 값과 그 값의 확률을 계산한 표

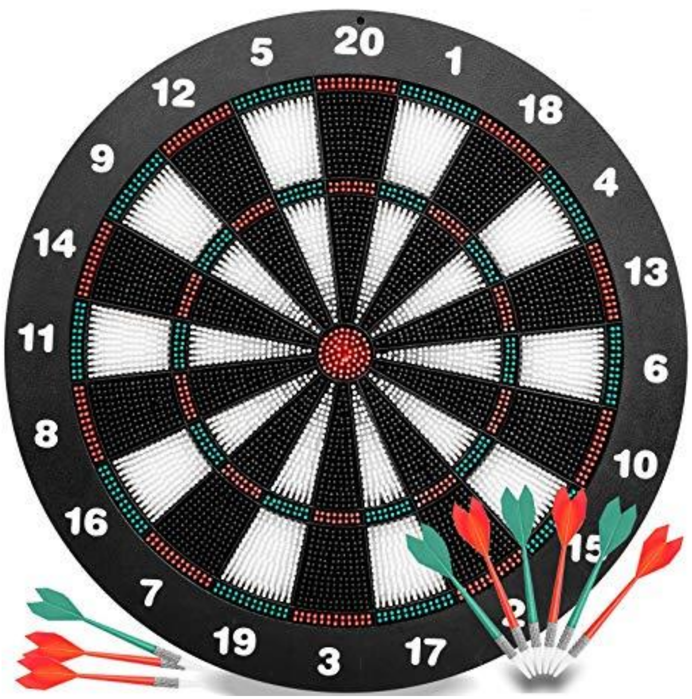
x	0	1	2
$\Pr(X = x)$	1/4	1/2	1/4



확률 분포

[확률 변수]

확률 변수의 값이 연속인 경우: 다트 위의 임의의 한 점



4번 구획에 들어갈 확률은?

- 면적의 비로 계산
- 전체 면적은 πr^2, 4번 구획의 면적은 $\pi r^2/20$

$$\frac{\pi r^2/20}{\pi r^2} = \frac{1}{20}$$

확률 분포

[**확률 밀도 함수**]

$$P(A) = \Pr(X \in A), \forall A \subset \chi$$

모든 A에 대하여 해당 확률 값을 계산하여 확률 분포표 생성?
어떤 A가 정의된다 하더라도 이를 계산할 수 있는 함수가 필요
➔ 확률 밀도 함수(probability density function)

확률 분포

[확률 밀도 함수 / 이산형 확률 밀도 함수]

확률 변수 공간이 이산형인 경우

$$f(x) > 0, \forall x \in \mathcal{X}$$

$$\sum_{\mathcal{X}} f(x) = 1$$

$$P(A) = \Pr(\mathrm{X} \in \mathrm{A}) = \sum_{A} f(x)$$

확률 분포

[확률 밀도 함수 / 이산형 확률 밀도 함수]

확률 변수 공간이 이산형인 경우

$$f(x) > 0, \forall x \in \mathcal{X}$$

$$\mathcal{X} = \{0, 1, 2, 3, 4\}$$
$$A = \{0, 1\}$$
$$f(x) = \frac{4!}{x!\,(4-x)!}\left(\frac{1}{2}\right)^4$$

동전은 4번 던질 때 앞면이 나오는 개수의 분포

$$\sum\nolimits_{\{0,1,2,3,4\}} f(x) = 1$$

$$(a+b)^n = \sum_x \binom{n}{x} b^x a^{n-x}$$
$$a = b = \frac{1}{2}, n = 4$$

$$P(A) = \sum\nolimits_{\{0,1\}} f(x)$$
$$= \frac{4!}{0!\,4!}\left(\frac{1}{2}\right)^4 + \frac{4!}{1!\,3!}\left(\frac{1}{2}\right)^4$$
$$= \frac{5}{16}$$

확률 분포

[**확률 밀도 함수 / 이산형 확률 밀도 함수**]

예제)
앞의 예에서 $A = \{0, 1, 2\}$의 확률을 계산해보자.
프로그램 또는 직접 계산

확률 분포

[확률 밀도 함수 / 이산형 확률 밀도 함수]

예제)
앞의 예에서 $A = \{0, 1, 2\}$의 확률을 계산해보자

[직접 계산]
대칭임을 이용하여 $1 - \frac{5}{16} = \frac{11}{16}$

확률 분포

[**확률 밀도 함수 / 이산형 확률 밀도 함수**]

예제)
앞의 예에서 $A = \{0, 1, 2\}$의 확률을 계산해보자

[프로그램 계산]

dbinom(x, size, prob) 이용

확률 분포

[확률 밀도 함수 / 연속형 확률 밀도 함수]

확률 변수 공간이 연속형인 경우

$f(x) > 0, \forall x \in \mathcal{X}$

$$\int_{\mathcal{X}} f(x)dx = 1$$

$$P(A) = \mathrm{Pr}(\mathrm{X} \in \mathrm{A}) = \int_{A} f(x)dx$$

확률 분포

[확률 밀도 함수 / 연속형 확률 밀도 함수]

확률 변수 공간이 연속형인 경우

$\mathcal{X} = \{0 < x < \infty\}, \qquad A = \{0 < x < 1\}, \qquad f(x) = e^{-x}$

$$\int_{\mathcal{X}} f(x)dx = 1\,?$$

$$P(A) = \Pr(\mathrm{X} \in \mathrm{A}) = ?$$

확률 분포

[확률 밀도 함수 / 연속형 확률 밀도 함수]

확률 변수 공간이 연속형인 경우

$$\mathcal{X} = \{0 < x < \infty\}, \qquad A = \{0 < x < 1\}, \qquad f(x) = e^{-x}$$

$$P(A) = \int_0^1 e^{-x}dx = [-e^{-x}]_0^1 = -e^{-1} - (-1) = 1 - e^{-1}$$

확률 분포

[**분포 함수(distribution function)**]

- $A = (-\infty, x]$인 경우에 $P(A) = \Pr(A) = F(x)$를 분포 함수라고 정의
- x 값에만 의존하는 함수

$$F(x) = \sum_{i \le x} f(i)$$

$$F(x) = \int_{-\infty}^{x} f(z) dz$$

확률 분포

[**분포 함수(distribution function)**]

- $A = (-\infty, x]$인 경우에 $P(A) = \Pr(A) = F(x)$를 분포 함수라고 정의
- x 값에만 의존하는 함수

$$0 \le F(x) \le 1 \because 0 \le \Pr(X \le x) \le 1$$

$$x_1 \le x_2 \Rightarrow F(x_1) \le F(x_2)$$

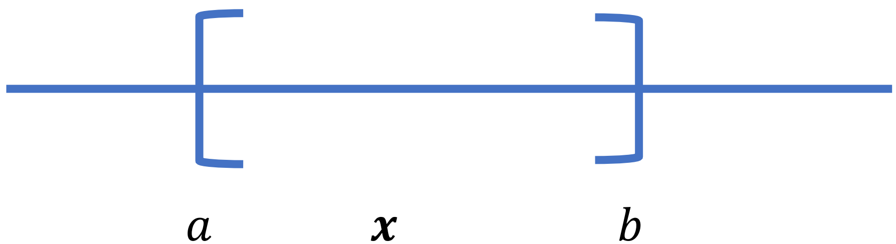
$$F(\infty) = 1, F(-\infty) = 0$$

$$\Pr(a < x < b) = F(b) - F(a)$$

확률 분포

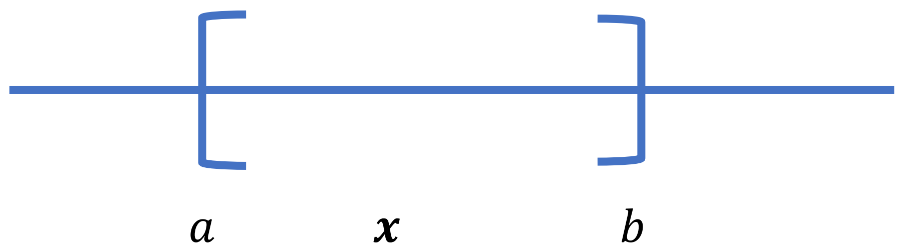
[분포 함수(distribution function)]

- $\chi = [a, b]$
- $X(x) = x$
- $[a, b]$ 에서 한 점을 선택하는 시행
- $F(x) = ?$



확률 분포

[분포 함수(distribution function)]



$$F(x) = \Pr([a, x]) = c(x - a)$$: 길이에 비례

$$F(b) = 1 \Rightarrow c(b - a) = 1 \Rightarrow c = \frac{1}{b - a}$$

$$\therefore F(x) = \frac{x - a}{b - a}$$

구간 $[a, b]$에서 균등 분포(uniform distribution)

확률 분포

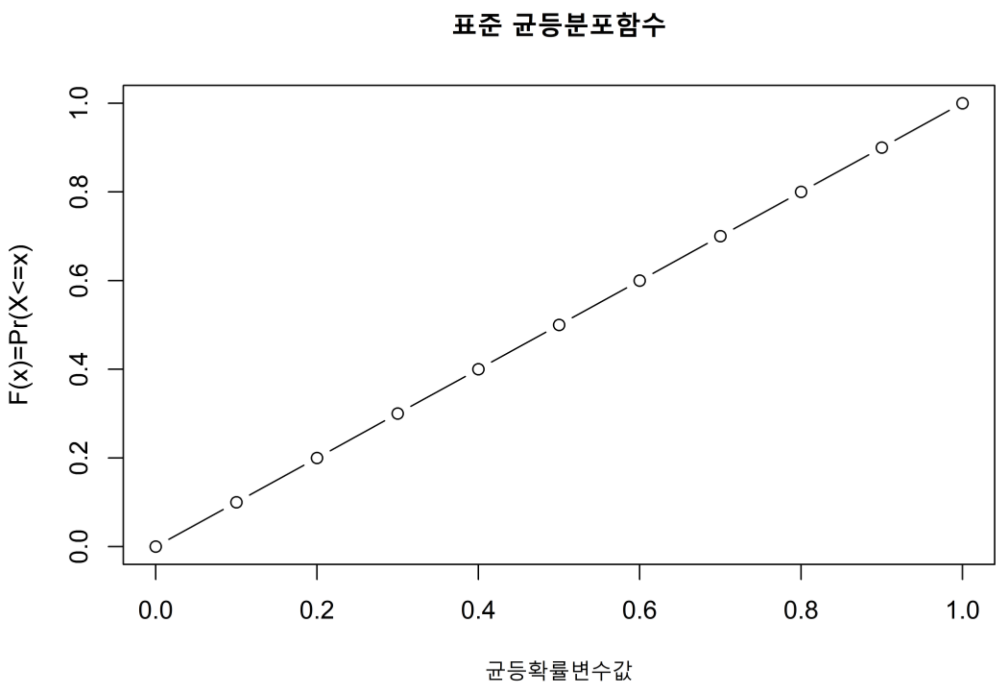
[**분포 함수(distribution function)**]

예제)

- 구간 [0, 1]에서 정의된 균등 분포의 분포 함수를 정의하세요
- 분포 함수를 x값에 따른 그래프를 그리세요
- punif 함수 이용

확률 분포

[분포 함수(distribution function)]



확률 분포

[**기대값(expectation)]**

- 확률변수 X 에 대한 기대값
- $E(X) = \sum_x xf(x)$
- X가 가질 수 있는 값이 $x_1, \ldots, x_n$ 일 때 이 값들이 가중 평균(weight average): $x_1 f(x_1) + \cdots + x_n f(x_n)$
- $\boldsymbol{\mu} = E(X)$

확률 분포

[기대값(expectation)]

- 확률변수 $(X-\mu)^2$에 대한 기대값: 분산
- $E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x)$
- 평균과의 편차 제곱에 대한 가중 평균
- $\boldsymbol{\sigma^2} = E(X-\mu)^2$
- σ: 분산의 양의 제곱근(표준편차)

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

확률 분포

[기대값(expectation)]

- 표준편차의 의미

예제)

$X \sim U(-1,1), Y \sim U(-2,2)$ 일 때

1) 각각의 평균을 구하세요
2) 각각의 표준편차를 구하세요

확률 분포

[기대값(expectation)]

- 표준편차의 의미

예제)

$X \sim U(-1,1), Y \sim U(-2,2)$ 일 때

$$E(X) = \int_{-1}^{1} x \times \frac{1}{2} dx = \left[\frac{1}{4}x^2\right]_{-1}^{1} = 0$$

$$E(X^2) = \int_{-1}^{1} x^2 \times \frac{1}{2} dx = \left[\frac{1}{6}x^3\right]_{-1}^{1} = \frac{1}{3}$$

$$\sigma^2 = \left(\frac{1}{3}\right) - (0)^2 = \frac{1}{3} \qquad \sigma = \frac{1}{\sqrt{3}} \qquad \sigma(Y) = 2/\sqrt{3}$$

확률 분포

[기대값(expectation)]

- 표준편차의 의미

예제)

$X \sim U(a, b)$ 일 때 평균과 분산은?

확률 분포

[기대값(expectation)]

- 표준편차의 의미

예제)

$X \sim U(a, b)$ 일 때 평균과 분산은?

$$\mu = \frac{a+b}{2}$$

$$\sigma^2 = \frac{1}{12}(b-a)^2$$

확률 분포

[기대값(expectation)]

- 적률 생성 함수(mgf: moment generating function)
- $E(X^n)$의 값을 구할 수 있는 함수

$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

$$M(0) = 1$$
$$\frac{d}{dt}M(t) = M'(t) = \int_{-\infty}^{\infty} xe^{tx} f(x)dx \; M'(0) = \int_{-\infty}^{\infty} xf(x)dx = E(X)$$
$$M''(0) = E(X^2)$$
$$M'''(0) = E(X^3), \ldots$$

확률 분포

[기대값(expectation)]

- 적률 생성 함수(mgf: moment generating function)
- $E(X^n)$의 값을 구할 수 있는 함수

$X \sim N(\mu, \sigma^2)$ 일 때, $M(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$

$$M'(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)(\mu + \sigma^2 t), M(0) = \mu$$
$$M''(t) = ?$$

확률 분포

[기대값(expectation)]

- 적률 생성 함수(mgf: moment generating function)
- $E(X^n)$의 값을 구할 수 있는 함수

$X \sim N(\mu, \sigma^2)$ 일 때, $M(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$

$$M'(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)(\mu + \sigma^2 t), M(0) = \mu$$
$$M''(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)(\mu + \sigma^2 t)(\mu + \sigma^2 t) + exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)(\sigma^2)$$
$$M''(0) = E(X^2) = \mu^2 + \sigma^2$$

확률 분포

[기대값(expectation)]

- 적률 생성 함수(mgf: moment generating function)
- $E(X^n)$의 값을 구할 수 있는 함수

$X \sim N(\mu, \sigma^2)$ 일 때, $M(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$

$$M'(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)(\mu + \sigma^2 t), M(0) = \mu$$
$$M''(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)(\mu + \sigma^2 t)(\mu + \sigma^2 t) + exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)(\sigma^2)$$
$$M''(0) = E(X^2) = \mu^2 + \sigma^2$$