인공지능 수학: 통계학
4. 몇 가지 확률분포

윤상민

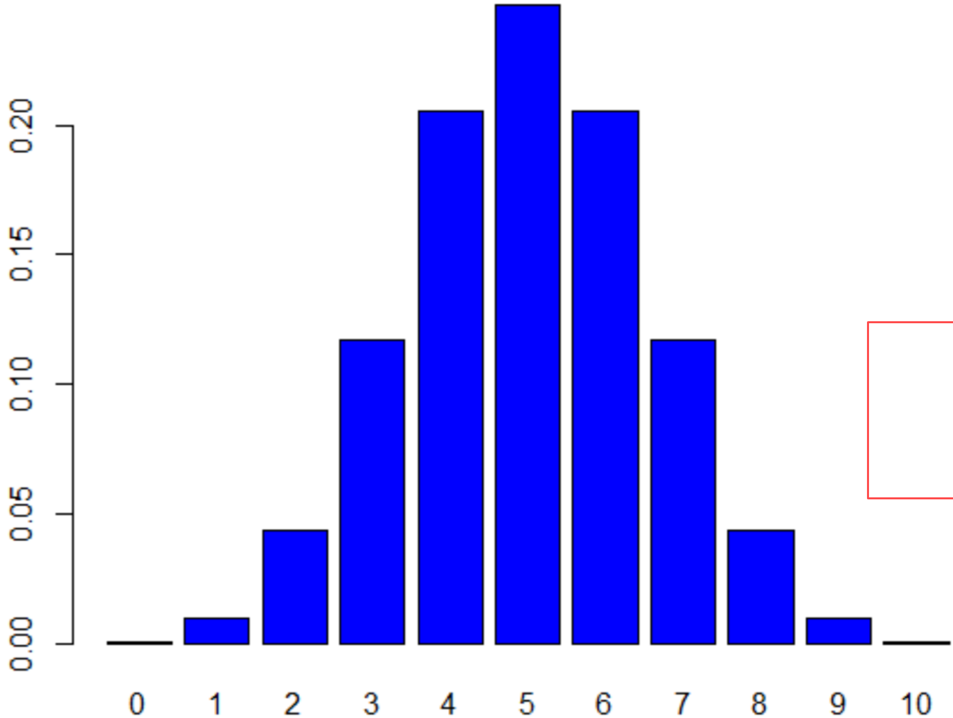
E-mail : smyoon@kookmin.ac.kr
Office : 02-910-4645

이항분포

- 베르누이 시행 (Bernoulli trial)
  - 정확하게 **2개의 결과**만을 가지는 실험
    - 예) 동전 던지기
  - 보통 성공과 실패로 결과를 구분
  - 성공의 확률: p
- 확률변수 X
  - **n번의 베르누이** 시행에서 성공의 횟수
  - 이항확률변수라고 함
- 이항분포 (binomial distribution)
  - 이항확률변수의 확률분포

이항분포

- 이항 확률변수 X의 확률분포
  - $f(x) = P[X = x] = \boldsymbol{\binom{n}{x} p^x (1-p)^{n-x}}$



$p = 0.5$
$n = 10$

이항분포

- 어떤 랜덤박스의 뽑기 성공 확률이 0.2이다.
- 3개를 뽑았을 때, 적어도 하나 이상의 성공이 발생할 확률은?
  - $P[X \geq 1] = 1 - P[X = 0] = 1 - \binom{3}{0}(0.2)^0(1 - 0.2)^{3-0} = 1 - 0.512 = 0.488$

```
>>> from scipy import stats
>>> 1 - stats.binom.cdf(0, n=3, p=0.2)
0.48799999999999999
```

이항분포

- 평균
  - $E(X) = np$
- 분산
  - $\mathrm{Var}(X) = np(1-p)$
- 표준편차
  - $\mathrm{SD}(X) = \sqrt{np(1-p)}$

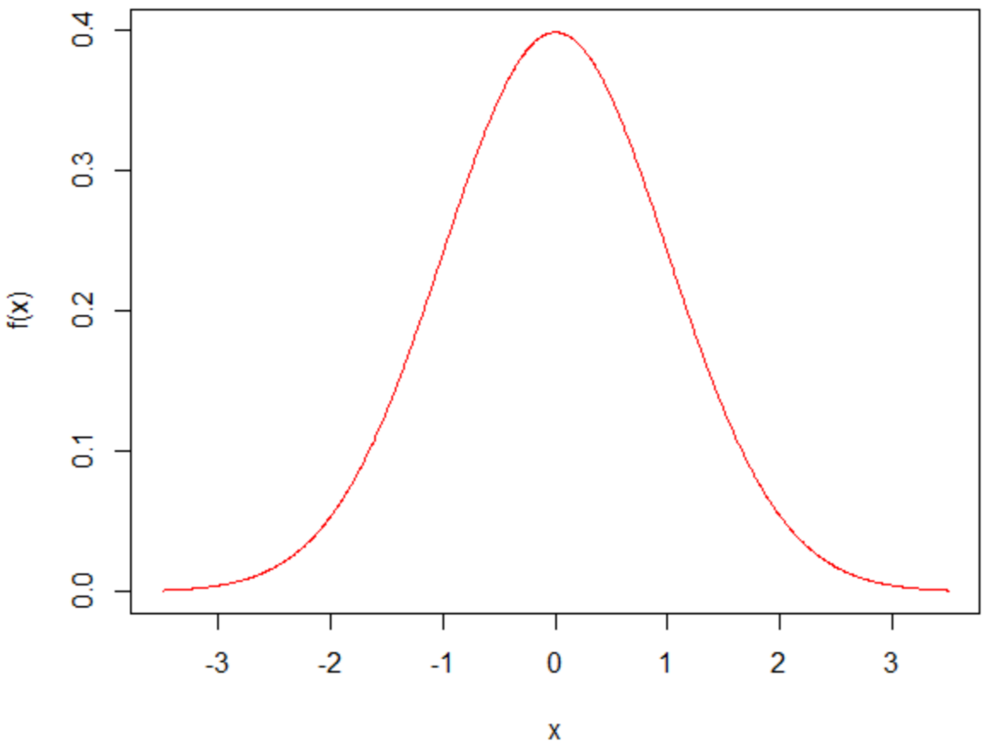
```
>>> stats.binom.stats(n=3, p=0.2)
(array(0.6), array(0.48))
```

정규분포

- 연속확률 변수의 확률 분포
  - 확률밀도함수 (probability density function)
    - $f(x)$
  - $P[a \leq X \leq b] = \int_a^b f(x)dx$
  - 즉 그래프 아래 부분의 넓이가 확률이 됨

정규분포

- 정규분포의 확률밀도함수
  - $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$



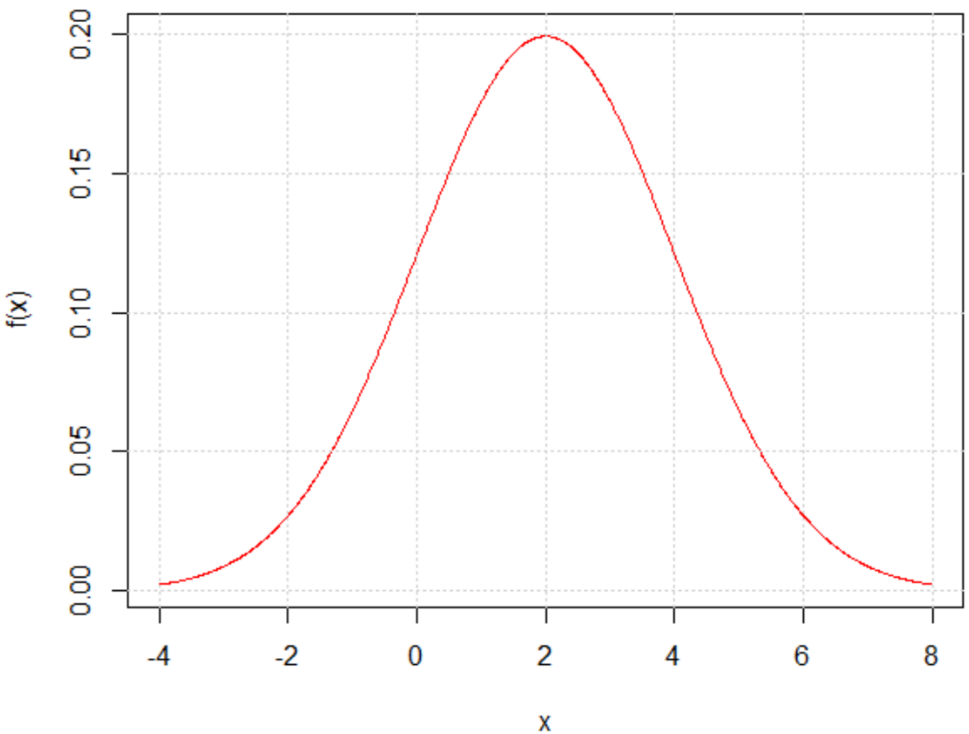
$$\mu = 0$$
$$\sigma = 1$$

$$X \sim N(\mu, \sigma^2)$$

정규분포

- 정규분포의 확률밀도함수
  - $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

$$\mu = 2$$
$$\sigma = 2$$



정규분포

- 표준정규확률변수
  - standard normal random variable

$$Z = \frac{X - \mu}{\sigma}$$

- 표준정규분포
  - standard normal distribution
  - $Z \sim N(0,1)$
  - 표준 정규분포표
    - $P[Z \leq z]$

정규분포

- $X \sim N(4, 3^2)$
  - $P[X \leq 4] = ?$

$$= P\left[\frac{X-\mu}{\sigma} \leq \frac{4-\mu}{\sigma}\right] = P\left[Z \leq \frac{4-4}{3}\right] = P[Z \leq 0]$$
$$= 0.5$$

```
>>> stats.norm.cdf(4, loc=4, scale=3)
0.5
```

정규분포

- $X \sim N(4, 3^2)$
  - $P[4 \le X \le 7] =?$

$$P[X \le 7] - P[X < 4] = P\left[Z \le \frac{7-4}{3}\right] - P[Z < 0]$$
$$= P[Z \le 1] - P[Z < 0] = 0.8413 - 0.5 = 0.3413$$

```
>>> stats.norm.cdf(7, loc=4, scale=3)-
stats.norm.cdf(4, loc=4, scale=3)
0.3413447460685429
```

정규분포

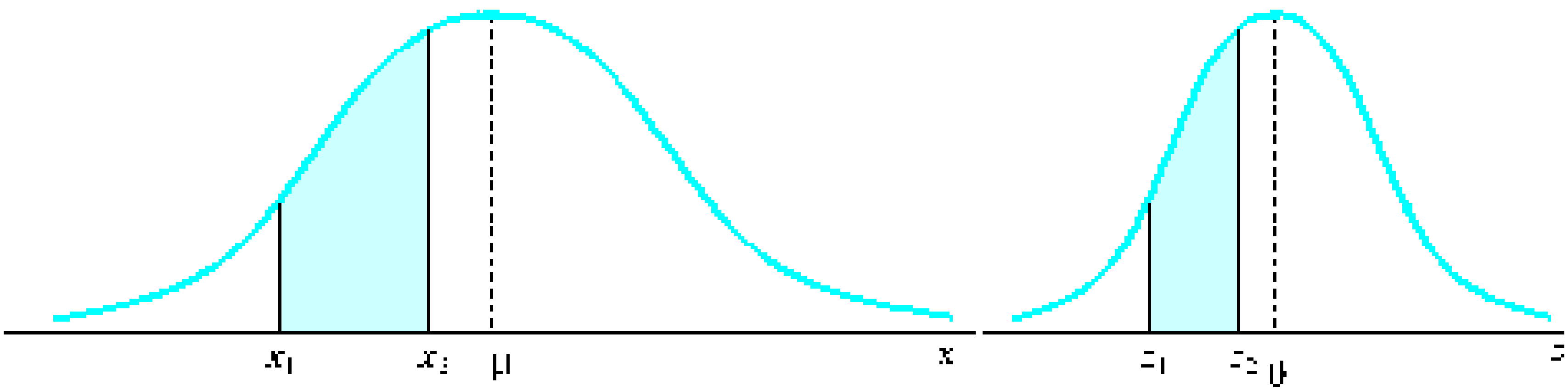
- 어떤 종목의 주가는 전날 종가를 평균으로 하고, 표준편차가 50인 정규분포를 따른다고 한다.
- 오늘 종가가 1,000원일 때, 내일 주가가 1,100원이상이 될 확률은?
- $P[X \geq 1100] = P\left[Z \geq \frac{1100-1000}{50}\right] = P[Z \geq 2] = 1 - P[Z < 2] = 1 - 0.9772 = 0.0228$

```
>>> 1 - stats.norm.cdf(1100,
loc=1000, scale=50)
0.022750131948179210
```

표준정규분포

- 임의의 평균 μ 와 분산 을 갖는 모든 정규확률변수를 평균이 0, 분산이 1로 표준화한 분포

$$z_1 = \frac{x_1 - \mu}{\sigma}, \quad z_2 = \frac{x_2 - \mu}{\sigma}$$



표준정규분포표

- 표준정규분포표에서 $P(Z<k)=0.3264$ 을 만족하는 *k* 값을 구하고, $P(k<Z<-0.18)=0.4197$ 을 만족하는 값을 구하시오.

표준정규분포표에서 $1-0.3264=0.6736$을 찾아서 반대로 z값을 읽어 보면 -0.45이므로 $k=-0.45$가 된다. 그리고 $P(k<Z<-0.18)=P(Z<-0.18)-P(Z<k)=0.4197$이고 $P(Z<-0.18)=0.4286$이므로 $P(Z<k)=0.0089$가 되어야 한다. 따라서 표에서 $1-0.0089=0.9911$을 찾아서 반대로 z값을 읽으면 $k=-2.37$이 된다.

표준정규분포표

- 통계학 시험성적이 평균이 50, 분산이 100인 정규분포를 따른다고 할 때, 45점과 62점 사이의 점수를 받은 학생의 비율을 구하시오.

$x_1 = 45,\ x_2 = 62$에 대응하는 z값은 $z_1 = \dfrac{45-50}{10} = -0.5,\ \ z_2 = \dfrac{62-50}{10} = 1.2$가

되므로 $P(45 < X < 62) = P(-0.5 < Z < 1.2) = P(Z < 1.2) - P(Z < -0.5) =$

$0.8849 - 0.3085 = 0.5764$가 된다.

포아송 분포

- Poisson distribution
- 일정한 시간단위 또는 공간 단위에서 발생하는 이벤트의 수의 확률분포
  - 하루 동안 어떤 웹사이트를 방문하는 방문자의 수
  - 어떤 미용실에 한 시간동안 방문하는 손님의 수
  - 어떤 전기선 100미터당 발생하는 결함의 수
- 확률분포함수 (확률질량함수)
  - $P[X = x] = f(x) = \boldsymbol{\lambda}^{x} \frac{\boldsymbol{e}^{-\boldsymbol{\lambda}}}{\boldsymbol{x}!}, x = 0,1,2, \ldots$
  - 평균: λ
  - 분산: λ

포아송 분포

- 어느 웹사이트에 시간당 접속자 수는 평균이 3 ($\lambda = 3$)인 포아송 분포를 따른다고 한다.
- 앞으로 1시간 동안 접속자 수가 2명 이하일 확률은?
- $P[X \leq 2] =?$

포아송 분포

- $P[X \leq 2] = P[X = 0] + P[X = 1] + P[X = 2] = 3^0 \frac{e^{-3}}{0!} + 3^1 \frac{e^{-3}}{1!} + 3^2 \frac{e^{-3}}{2!} = 0.04998 + 0.14936 + 0.22404 = 0.42319$

```
>>> stats.poisson.cdf(2, mu=3)
0.42319008112684364
```