확률과 통계

섹션 - 2

강사 : James 쌤

이산확률 – Part 1

키포인트

- 확률변수와 확률분포.
- 이산확률과 연속확률.
- 모집단과 모수.
- 기대값 (모평균)의 법칙.
- 모분산의 법칙.

확률변수

- 확률변수 (random variable): 확률실험의 각 결과에 수치값을 부여하는 함수.
- 확률변수의 값은 하나의 수치로 나타낸 사건이다.

예). 동전을 한번 던지는 실험에서 앞면(H)이 나오면 1 뒷면(T)이 나오면 0.

확률변수의 유형

- 이산확률변수 (discrete random variable): 셀 수 있는 가지수의 값을 가지는 확률변수.

  예). 주사위를 던져서 나오는 눈의 수: $\{1, 2, 3, 4, 5, 6\}$

- 연속확률변수 (continuous random variable): 셀 수 없는 (무한대) 가지수의 값을 가지는 확률변수.

  예). 1년 연봉, 성인남성의 신장, 등.

확률분포

- 이산확률분포: 이산확률변수가 가지는 값과 이것의 **확률** 사이의 **대응 관계**.

  ⇒ 확률변수는 영문 대문자로 확률변수의 값은 영문 소문자로 표기한다.

  예). 확률변수 X, 확률변수의 값 x.

  ⇒ 확률변수 X의 값이 x일 확률은 $P(X = x)$ 또는 $P(x)$로 표기한다.

확률분포

- 연속확률밀도/연속확률분포: 이것을 사용하여 연속확률변수의 값이 특정 구간에 속할 확률을 나타낼 수 있다.

확률분포

- 연속확률밀도/연속확률분포: 이것을 사용하여 연속확률변수의 값이 특정 구간에 속할 확률을 나타낼 수 있다.

연속확률과 연속확률분포에 대해서는 차후 강의에서 자세히 다루기로 한다.

당분간 이산확률과 이산확률분포에 포커스를 두고 알아보도록 한다.

이산확률분포의 필수 조건

- 다음 조건이 충족되어야 한다.

$$0 \leq P(x) \leq 1$$

$$\sum_{all\ x_i} P(x_i) = 1$$

모집단과 모수

- 모집단 (population): 분석 대상 전체를 의미함. 실존 또는 개념적 존재.
- 모수 (parameter): 모집단을 특성을 나타낸다.

  예). 모평균, 모분산, 모표준편차, 등.
- 모집단의 특성을 묘사하기 위해서 확률분포함수를 사용할 수 있다.

  예). 모평균을 확률분포함수를 사용하여 계산한다.

이산확률을 따르는 모집단의 평균

- 모집단의 평균 "모평균"은 확률변수 X의 기대값 (expected value)이라고도 불리우며 $E[X]$와 같이 표기한다.

이산확률을 따르는 모집단의 평균

- 모평균은 확률변수 X의 기대값 (expected value)이라고도 불리우며 $E[X]$ 또는 μ와 같이 표기한다.
- 이산확률변수가 가질 수 있는 값들에 확률을 가중치로 곱해서 평균을 구한 것 (≃더한 것)이다.

$$\mu = E[X] = \sum_{all\ x} x\,P(x)$$

기대값 (모평균)의 법칙

- $E[c] = c$ $\Leftarrow$ c는 상수이다.
- $E[X + Y] = E[X] + E[Y]$
- $E[c\,X] = c\,E[X]$
- $E[X + c] = E[X] + c$

이산확률을 따르는 모집단의 분산 (모분산)

- 확률변수 X의 **모분산**은 $Var(X)$ 또는 σ^2 와 같이 표기한다.
- 모평균을 기준으로한 **편차의 제곱**에 확률을 가중치로 곱해서 평균을 구한 것 (≃더한 것).

$$\sigma^2 = Var(X) = \sum_{all\ x} (x-\mu)^2 P(x)$$

이산확률을 따르는 모집단의 분산 (모분산)

- 모분산의 간편 수식:

$$\sigma^2 = \left(\sum_{all\ x} x^2 \, P(x) \right) - \mu^2 = E[X^2] - (E[X])^2$$

- 모표준편차:

$$\sigma = \sqrt{\sigma^2}$$

모분산의 법칙

- $Var(c) = 0$ ⇐ 상수의 분산은 0이다.
- $Var(X + c) = Var(X)$
- $Var(c\,X) = c^{\mathbf{2}}\,Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2\,Cov(X, Y)$

$= Var(X) + Var(Y)$ ⇐ X와 Y가 서로 **독립**인 경우에만!

모평균과 모분산

문제: 동전 던지기 게임이 있다. 앞면이 나오면 $X = 100$을 받고 뒷면이 나오면 0을 받는다. 그런데 참가비용은 40이다. 수익의 평균과 표준편차는?

모평균과 모분산

문제: 동전 던지기 게임이 있다. 앞면이 나오면 $X = 100$을 받고 뒷면이 나오면 0을 받는다. 그런데 참가비용은 40이다. 수익의 평균과 표준편차는?

수익을 나타내는 확률변수 $= X - 40$

수익의 모평균 $= E[X-40]$

$= E[X] - 40$

$= 100 \times P(X = 100) + 0 \times P(X = 0) - 40$

$= 100 \times \frac{1}{2} - 40 = 50 - 40 = 10$

모평균과 모분산

문제: 동전 던지기 게임이 있다. 앞면이 나오면 $X = 100$을 받고 뒷면이 나오면 0을 받는다. 그런데 참가비용은 40이다. 수익의 평균과 표준편차는?

$$\text{수익의 모분산} = Var(X-40) = Var(X) = E[X^2] - (E[X])^2$$

$$= 100^2 \times P(X=100) + 0^2 \times P(X=0) - 50^2$$

$$E[X] = 50$$

$$= 10000 \times \frac{1}{2} - 2500 = 5000 - 2500 = 2500$$

$$\text{수익의 모표준편차} = \sqrt{2500} = 50$$ "리스크"

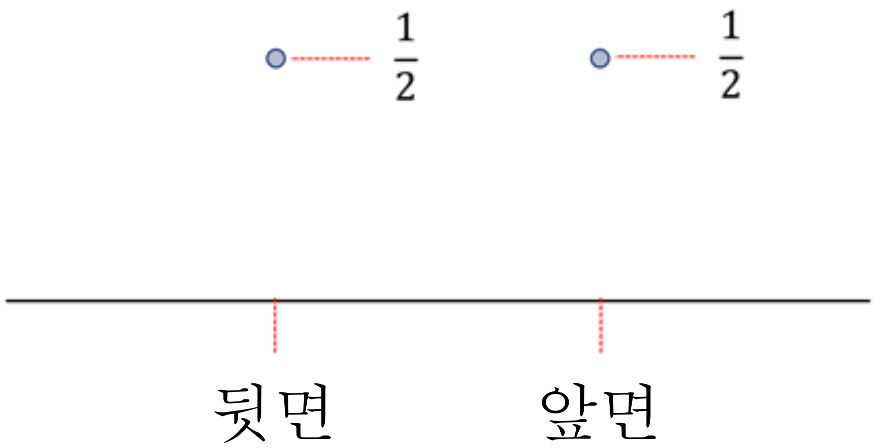
이산확률 – Part 2

키포인트

- 베르누이 확률분포 (Bernoulli).
- 이항 확률분포 (Binomial).

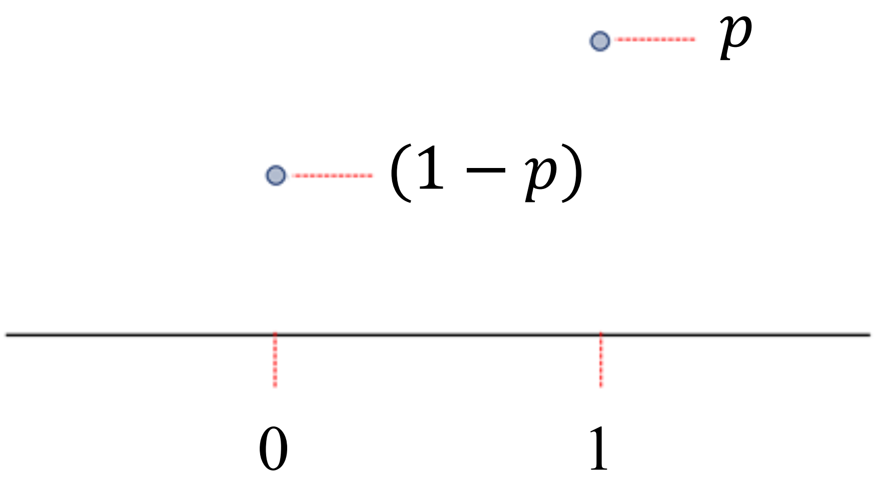
베르누이 확률분포

베르누이 확률분포



동전 던지기

베르누이 확률분포



베르누이 확률분포

p 는 0과 1 사이의 수치.

베르누이 확률분포

- 베르누이 시행에는 두개의 가능한 값이 있음:

  예). 1 또는 0, 동전의 앞면(H) 또는 뒷면(T), “성공” 또는 “실패”.

- 베르누이 시행에서 $P(X=1)=p$ 이고 $P(X=0)=(1-p)$이다.

- 베르누이 확률분포는 수학적으로 다음과 같이 표현이 가능하다.

$$P(x)=p^{x}(1-p)^{1-x}$$

- “확률변수 X가 베르누이 확률분포를 따른다.” $\Leftrightarrow$ $X \sim Ber(p)$

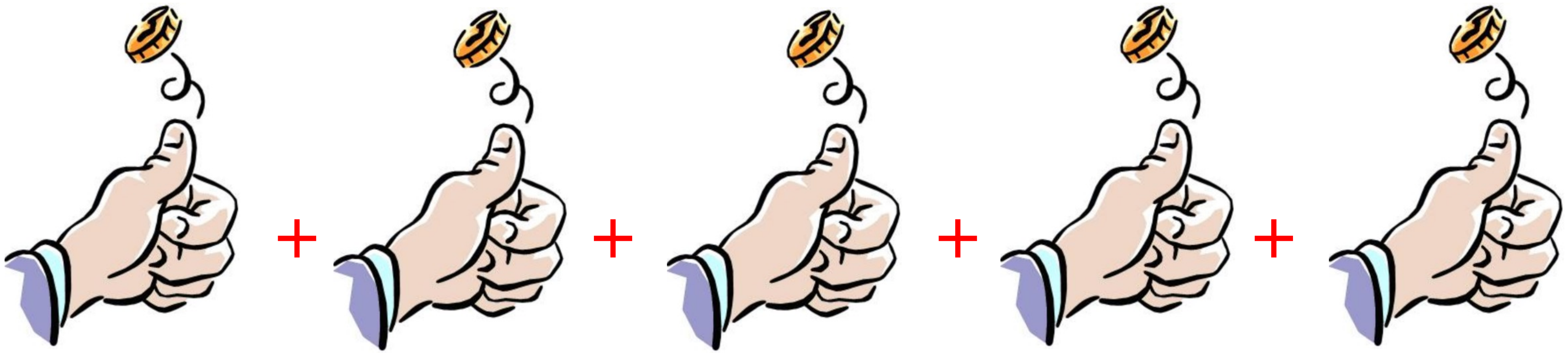
베르누이 확률분포

- 평균 $= 1\,P(1) \;+\; 0\,P(0) \;=\; P(1) \;=\; p$

- 분산 $= \{1^2\,P(1) + \;0^2\,P(0)\} - p^2 = \;P(1) - p^2 = p - p^2 = \;p\,(1-p)$

- 표준편차 $= \sqrt{p\,(1-p)}$

이항 확률분포



이항 확률분포

- 이항 확률변수 X_{bin}은 0 또는 1의 값을 갖는 n개의 독립적인 베르누이 확률변수 X_{Ber}를 더한 것.

$$X_{bin} = X_{Ber} + X_{Ber} + \cdots + X_{Ber}$$

← n 개 →

예). 동전 하나를 n 번 던져서 앞면(H)이 나온 횟수를 집계.

n회 시행하여 "성공"한 횟수를 더한다.

- "확률변수 X가 이항 확률분포를 따른다." ⇔ $X \sim Bin(n, p)$

이항 확률분포

- 이항 확률분포:

$$P(x) = \binom{n}{x} p^x (1-p)^{\,n-x}$$

- 소문자 p는 개개 베르누이 확률변수의 값이 1과 같을 확률.
- x는 0과 n 사이의 숫자 이어야 한다.

$$0 \le x \le n$$

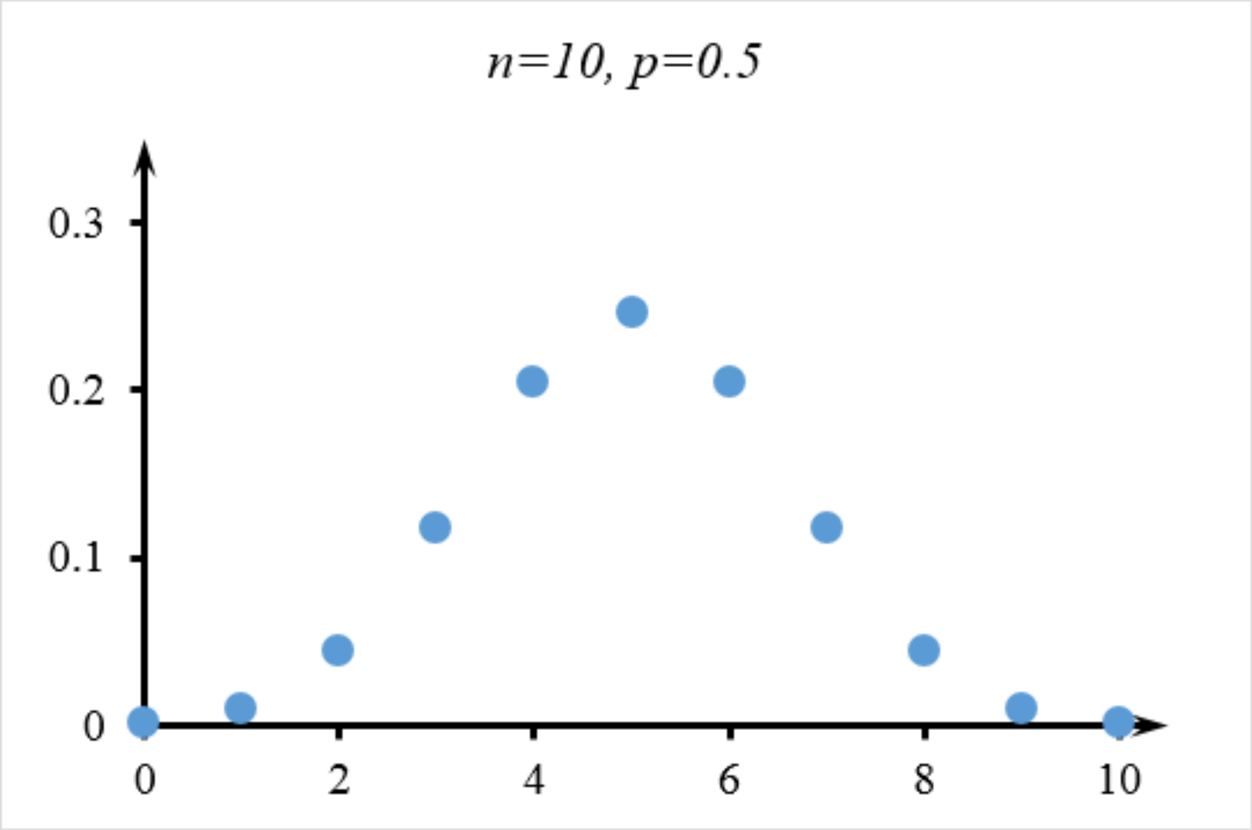
- $\binom{n}{x}$는 조합을 나타내고 $\frac{n!}{x!(n-x)!}$와 같이 계산한다.

이항 확률분포

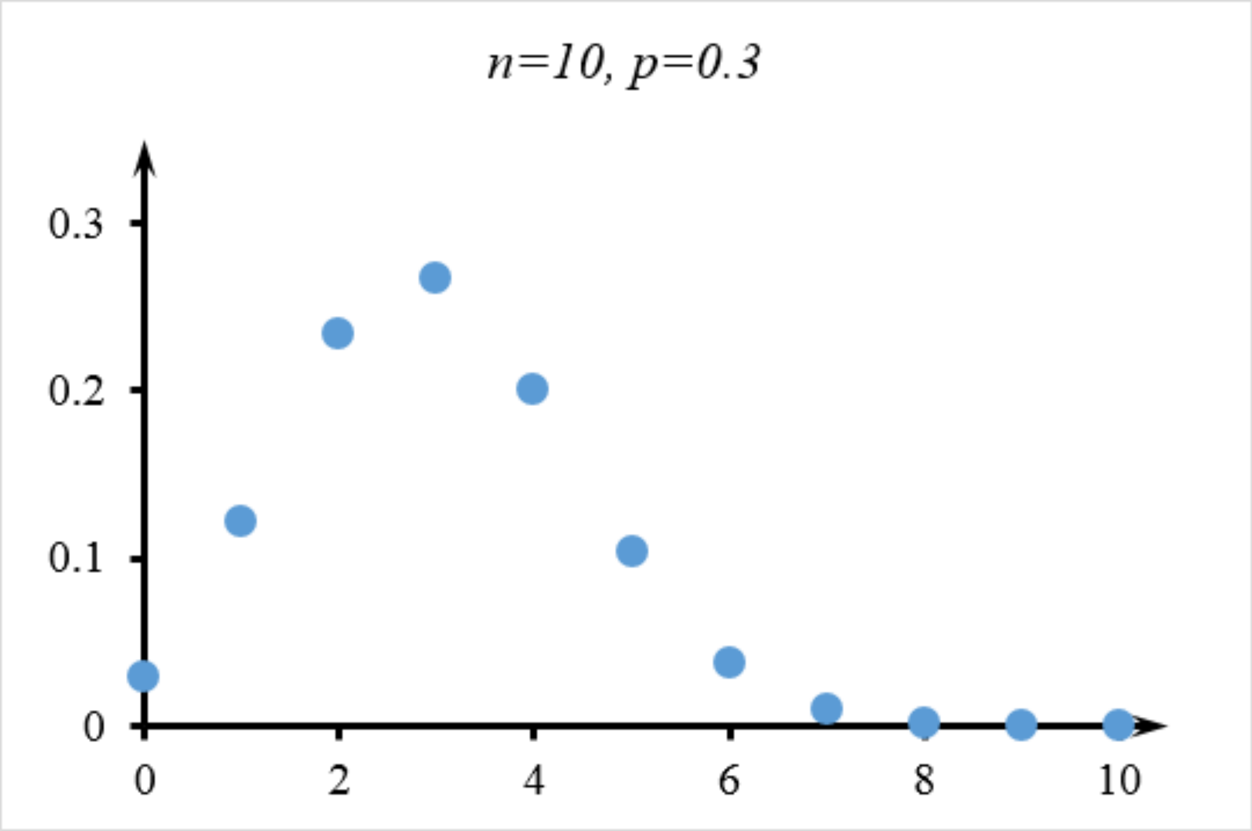
- 이항 확률변수의 정의를 그대로 적용하여 평균, 분산, 표준편차를 구하는 것이 이항 확률분포의 수식을 사용해서 구하는 것보다 **쉽다**.
- 평균 $= E[X_{bin}] = E[X_{Ber}] + \cdots + E[X_{Ber}] = nE[X_{Ber}] = n\,p$
- 분산 $= Var(X_{bin}) = Var(X_{Ber}) + \cdots + Var(X_{Ber})$ "서로 독립적"

  $= n\,Var(X_{Ber}) = n\,p\,(1-p)$
- 표준편차 $= \sqrt{n\,p\,(1-p)}$

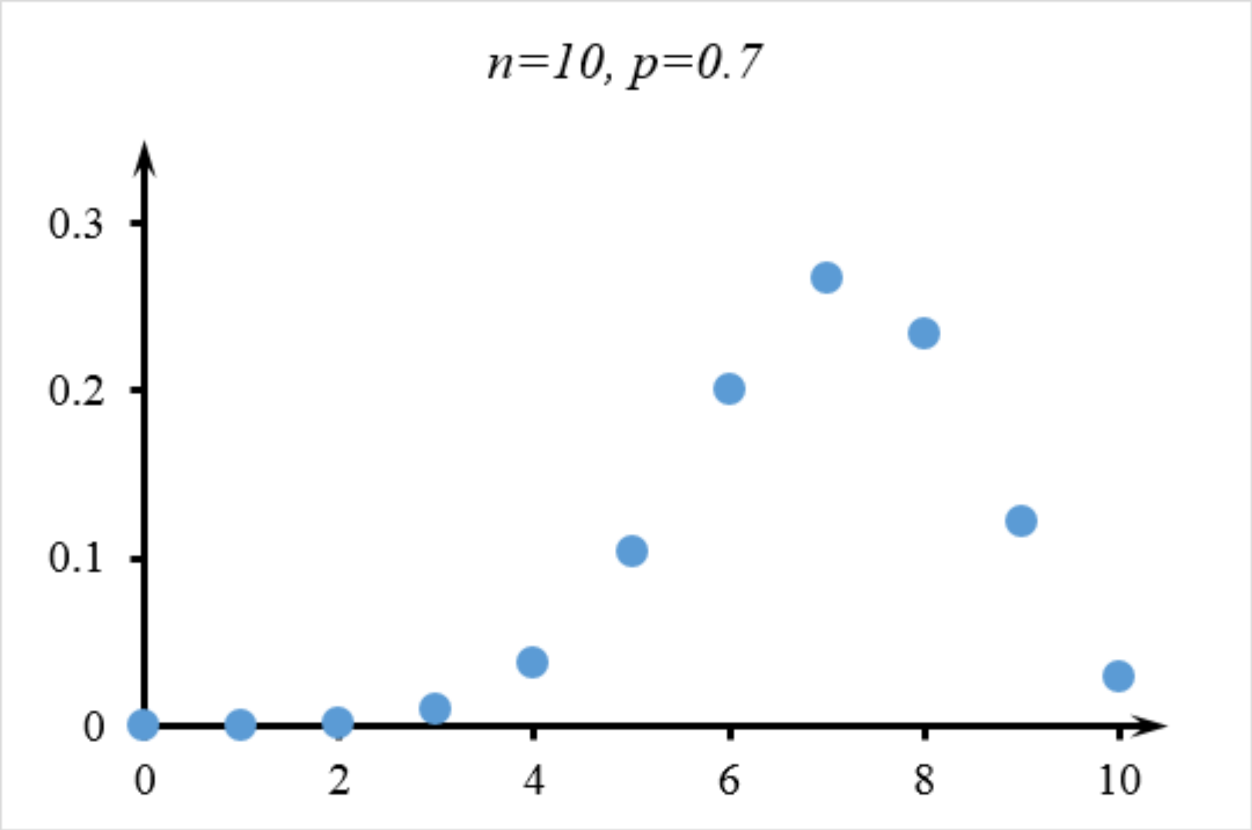
이항 확률분포



이항 확률분포



이항 확률분포



이항 확률변수의 합

- 서로 독립적인 확률변수 X와 Y가 다음과 같이 이항 확률분포를 따를 때,

$$\left.\begin{array}{l} X \sim Bin(n, p) \\ \\ Y \sim Bin(m, p) \end{array}\right\} X + Y \sim Bin(n + m, p)$$이다.

- 증명은 다음과 같이 매우 간단하다.

$$X + Y = \underbrace{\{X_{Ber} + X_{Ber} + \cdots + X_{Ber}\}}_{\leftarrow \quad n\text{개} \quad \rightarrow} + \underbrace{\{X_{Ber} + \cdots + X_{Ber}\}}_{\leftarrow \quad m\text{개} \quad \rightarrow}$$

$\Rightarrow$ 그러므로, $X + Y \sim Bin(n + m, p)$.

이항 확률분포

문제: 장마 기간이 왔다. 일일 강우확률은 30%라고 한다. 향후 5일간 날씨에 대한 질문에 답하시오.

1). 비가 한번도 내리지 않을 확률은?

이항 확률분포

문제: 장마 기간이 왔다. 일일 강우확률은 30%라고 한다. 향후 5일간 날씨에 대한 질문에 답하시오.

1). 비가 한번도 내리지 않을 확률은?

X가 강우 횟수를 나타내는 이항 확률변수라면 $P(X = 0)$이 구하고자 하는 답:

$$P(0) = \binom{5}{0} 0.3^0 (1 - 0.3)^{5-0}$$

$$= \frac{5!}{0!\,5!} 0.7^5 = 0.7^5 = 0.168$$ $\Rightarrow$ 대략 16.8%

이항 확률분포

문제: 장마 기간이 왔다. 일일 강우확률은 30%라고 한다. 향후 5일간 날씨에 대한 질문에 답하시오.

2). 비가 정확하게 2번 내릴 확률은?

이항 확률분포

문제: 장마 기간이 왔다. 일일 강우확률은 30%라고 한다. 향후 5일간 날씨에 대한 질문에 답하시오.

2). 비가 정확하게 2번 내릴 확률은?

$$P(2) = \binom{5}{2} 0.3^2 (1 - 0.3)^{5-2}$$

$$= \frac{5!}{2!3!} 0.3^2 \times 0.7^3 = 10 \times 0.3^2 \times 0.7^3 = 0.309 \Rightarrow \text{대략 } 30.9\%$$

이항 확률분포

문제: 장마 기간이 왔다. 일일 강우확률은 30%라고 한다. 향후 5일간 날씨에 대한 질문에 답하시오.

3). 비가 3회 이하 내릴 확률은?

이항 확률분포

문제: 장마 기간이 왔다. 일일 강우확률은 30%라고 한다. 향후 5일간 날씨에 대한 질문에 답하시오.

3). 비가 3회 이하 내릴 확률은?

$$P(X \leq 3) = P(0) + P(1) + P(2) + P(3)$$

$$= \binom{5}{0} 0.3^0 (1-0.3)^{\,5-0} + \binom{5}{1} 0.3^1 (1-0.3)^{\,5-1} + \binom{5}{2} 0.3^2 (1-0.3)^{\,5-2}$$

$$+ \binom{5}{3} 0.3^3 (1-0.3)^{\,5-3}$$

$$= 1 \times 1 \times 0.7^5 + 5 \times 0.3 \times 0.7^4 + 10 \times 0.3^2 \times 0.7^3 + 10 \times 0.3^3 \times 0.7^2$$

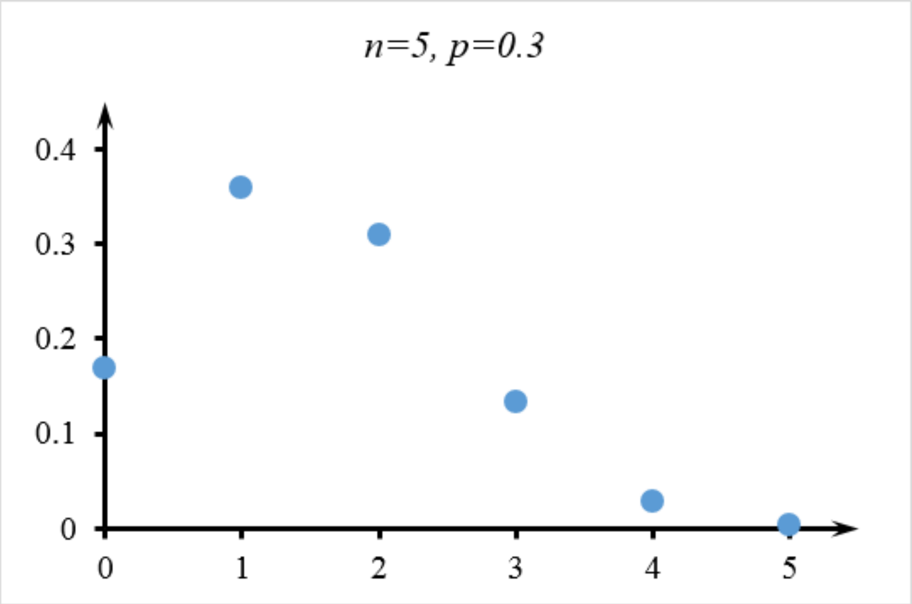
$$= 0.16807 + 0.36015 + 0.3087 + 0.1323$$

$$= 0.96922$$ $\Rightarrow$ 대략 96.9%

이항 확률분포

문제: 장마 기간이 왔다. 일일 강우확률은 30%라고 한다. 향후 5일간 날씨에 대한 질문에 답하시오.

3). 비가 3회 이하 내릴 확률은?



이산확률 – Part 3

키포인트

- 푸아송 확률분포 (Poisson).
- 이항과 푸아송의 관계.

푸아송 확률분포

- 푸아송 확률분포는 프랑스의 과학자 Simeon Poisson의 이름에서 유래한다.
- 일정 시간 또는 공간에서 발생하는 사건 (성공)의 횟수 (count)에 대한 이산 확률분포이다.

  예). 시간당 수신하는 이메일의 개수.

  특정 지역 (국가)의 년간 지진의 발생 횟수.

  초콜렛 칩 쿠키에 박힌 초콜렛 조각의 개수.

- "확률변수 X가 푸아송 확률분포를 따른다." $\Leftrightarrow X \sim Pois(\lambda)$

푸아송 확률분포

- 푸아송 확률분포:

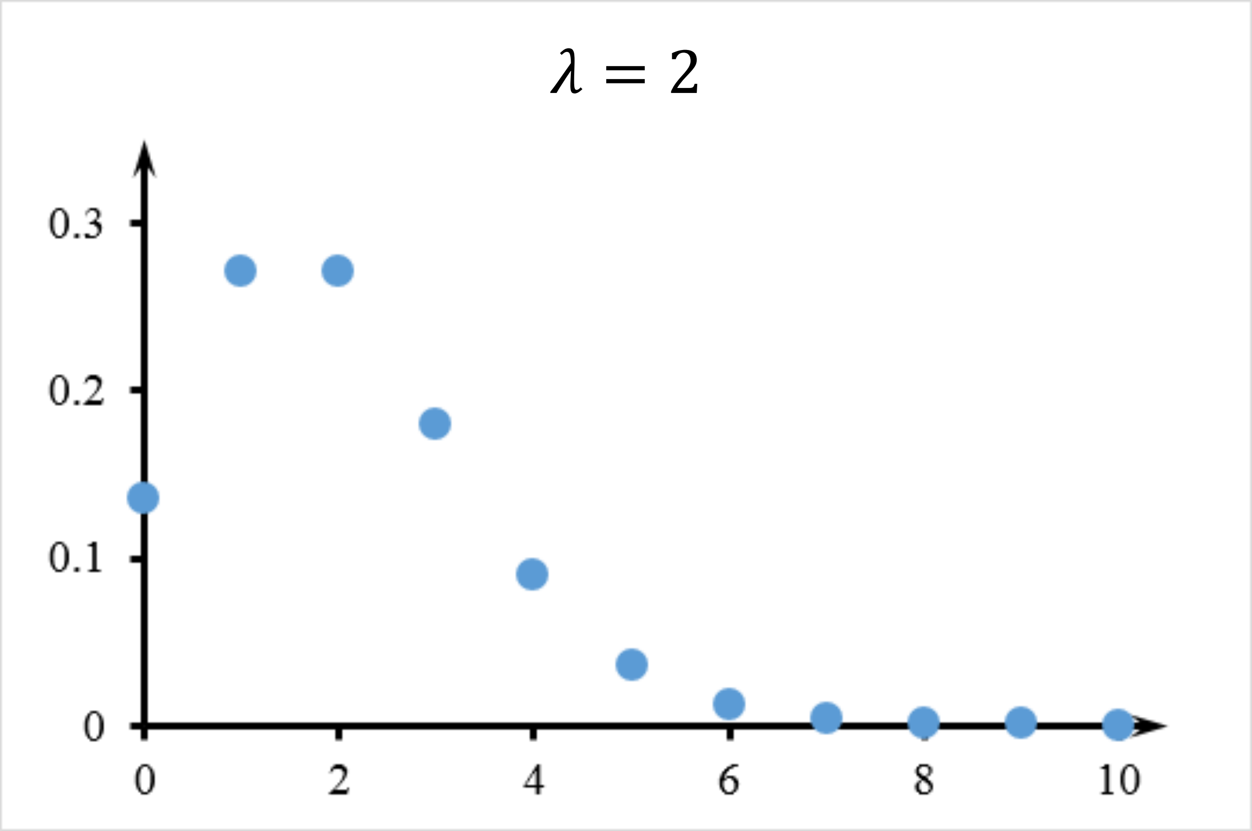
$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- λ는 유일한 파라미터이고 양의 수치여야 한다. 한 단위 시간 또는 공간에서 발생하는 사건 (성공) 횟수의 평균과 같다. 발생률 또는 성공률 이라고도 해석할 수 있다.
- x는 0 이상의 정수값 이어야 한다: $0 \leq x$

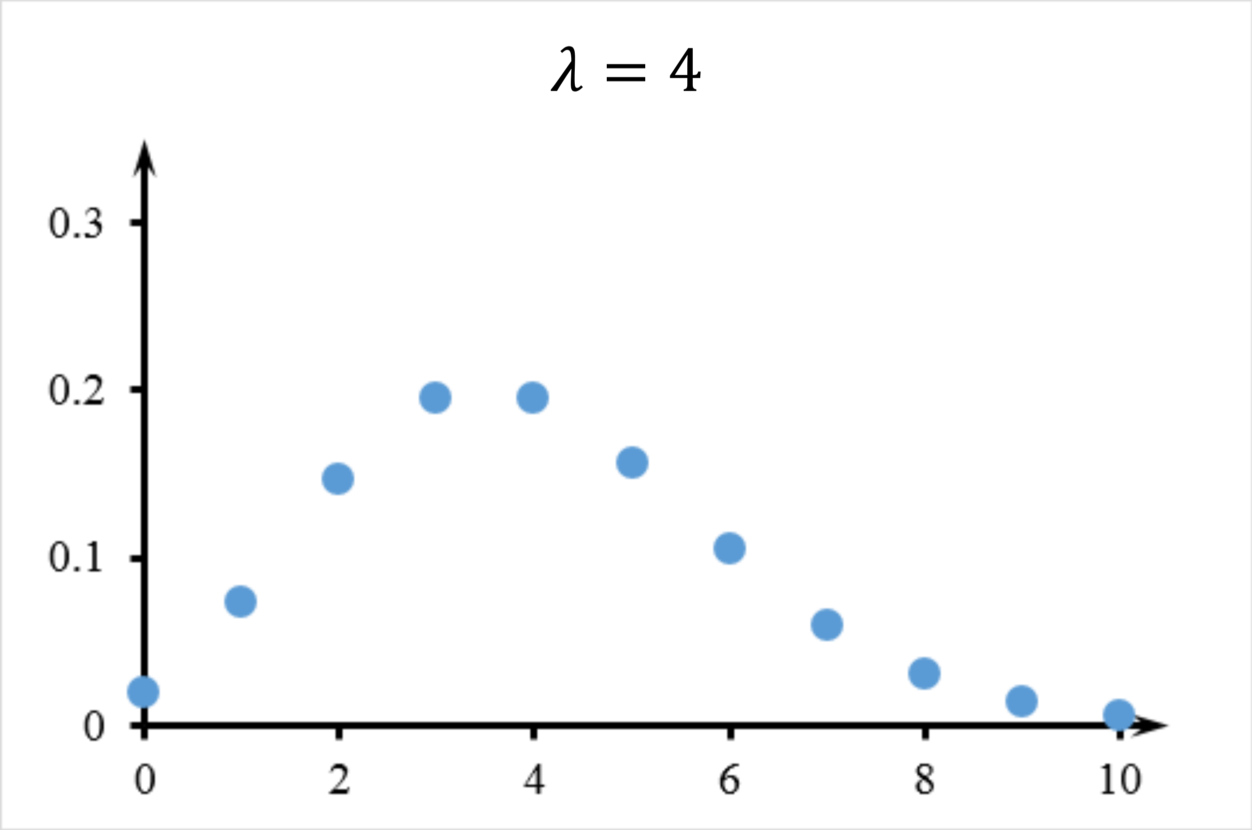
푸아송 확률분포

- 푸아송 확률변수의 평균과 분산은 일치한다.
- 평균 $= E[X_{Pois}] = \lambda$
- 분산 $= Var(X_{Pois}) = \lambda$
- 표준편차 $= \sqrt{\lambda}$

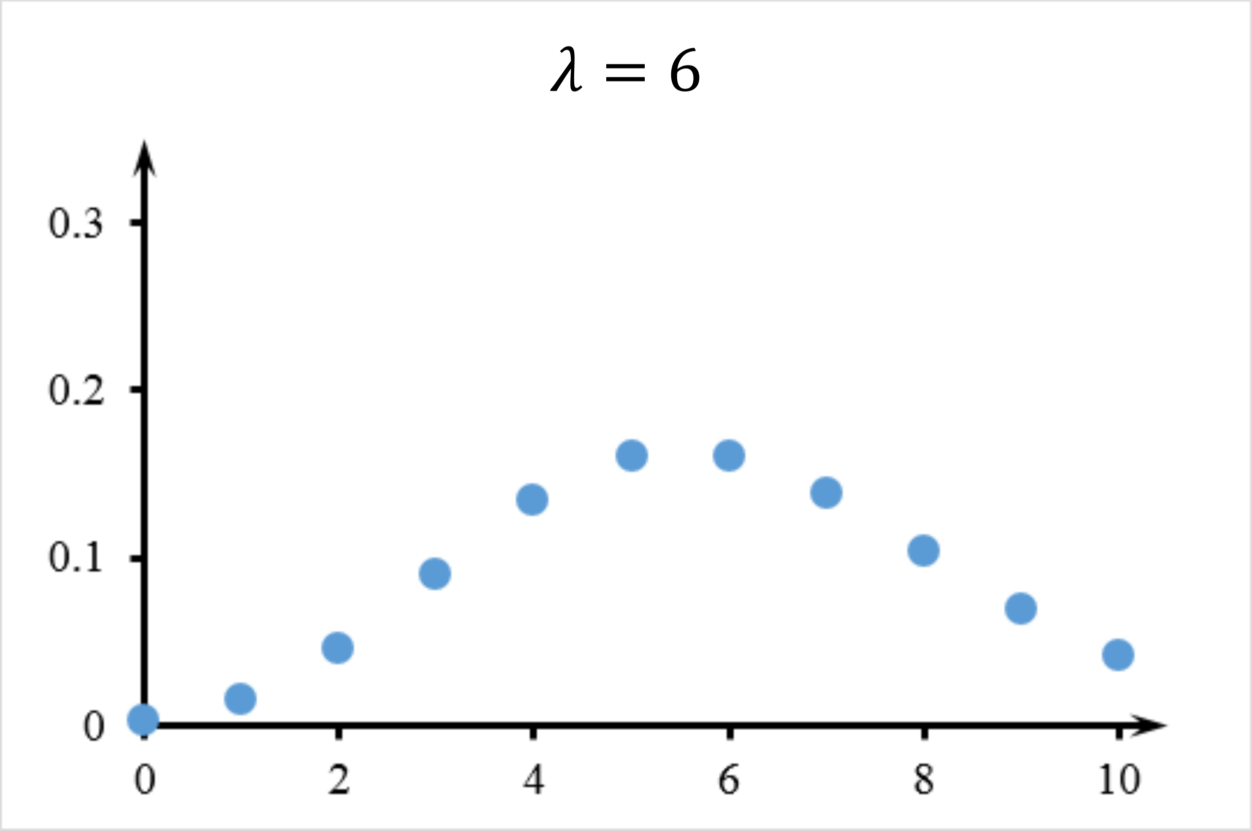
푸아송 확률분포



푸아송 확률분포



푸아송 확률분포



푸아송과 이항

- 푸아송과 이항은 이산확률을 나타낸다. 그리고,

  $\Rightarrow$ 이항 확률변수의 평균 $= n\,p$

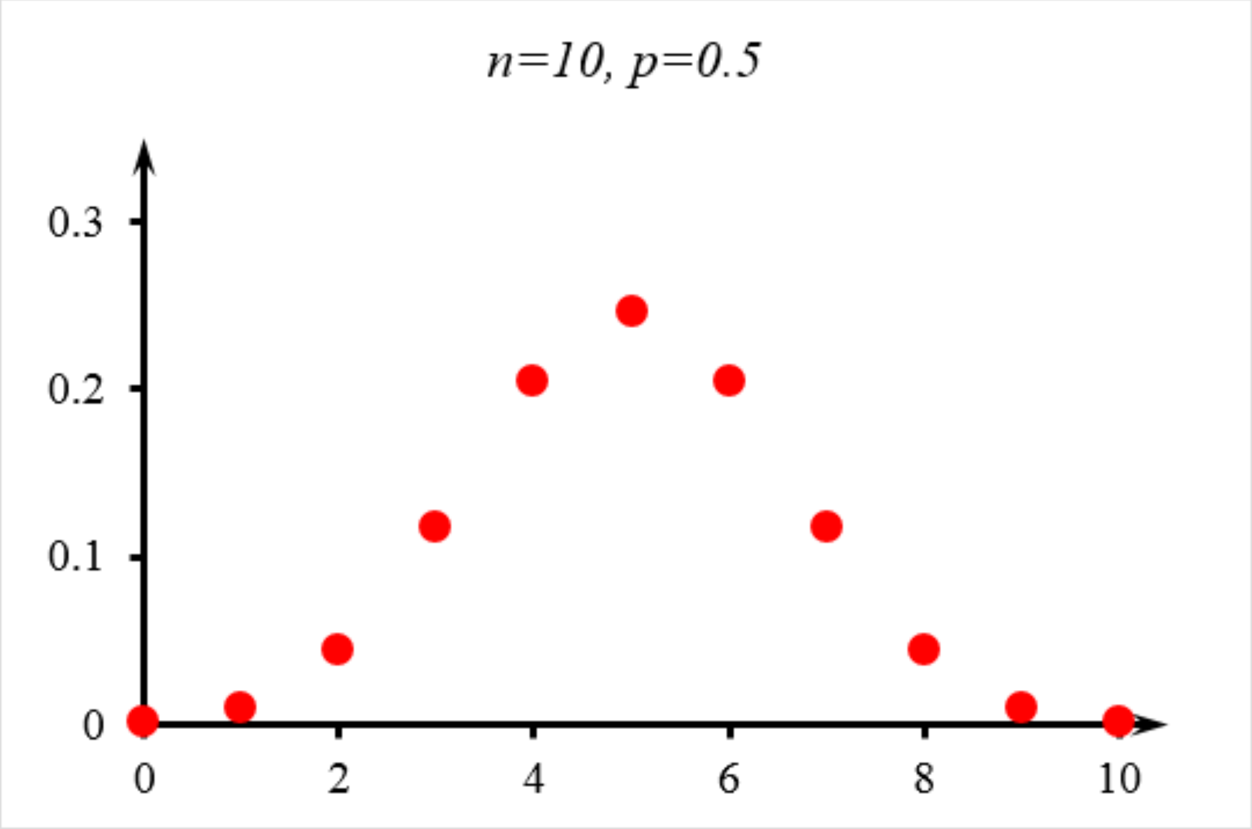
  $\Rightarrow$ 푸아송 확률변수의 평균 $= \lambda$

- 평균이 일치하는 조건 ($n\,p = \lambda$)을 유지하며 이항 확률의 n을 키움과 동시에 p를 줄이면, 이항 확률은 푸아송 확률에 수렴한다.

  $\Leftrightarrow p = \dfrac{\lambda}{n}$ 과 같이 선택한다는 의미.

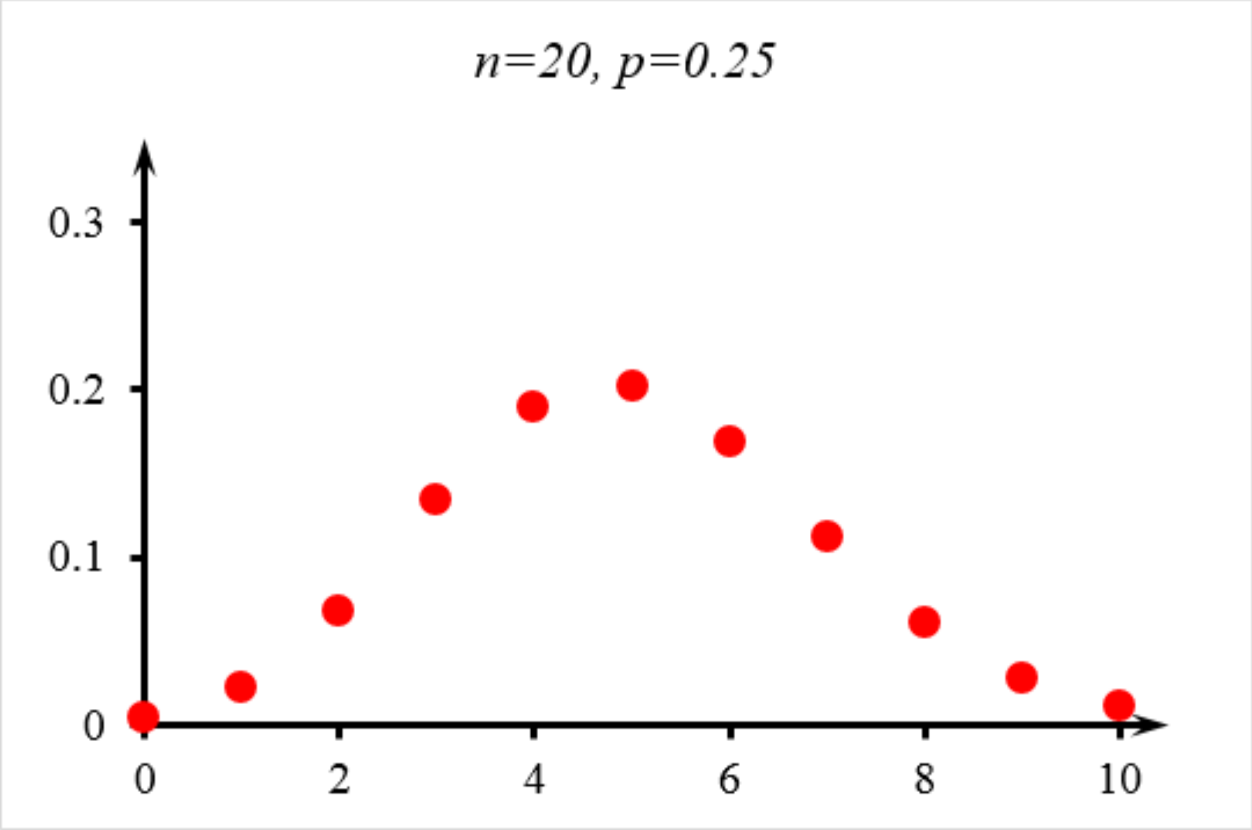
$$n\,p = \lambda$$

푸아송과 이항



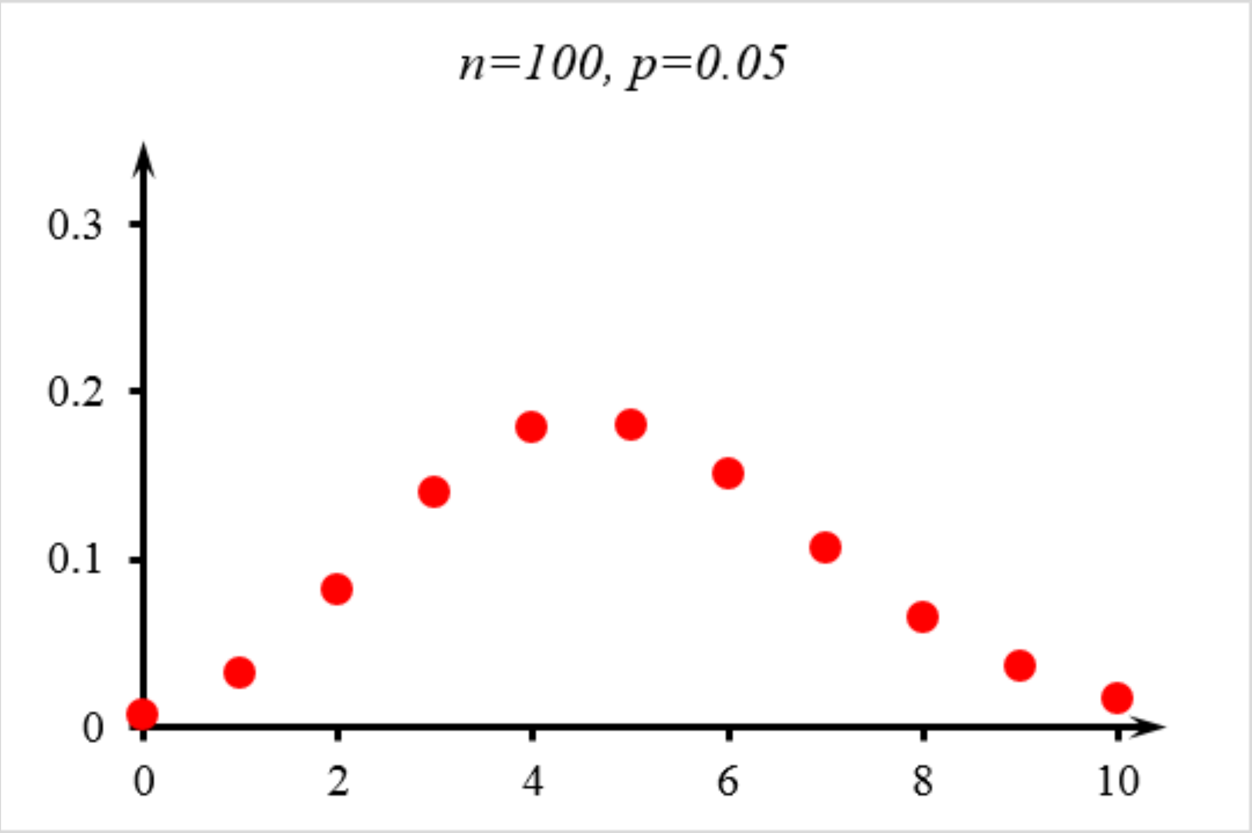
이항

푸아송과 이항



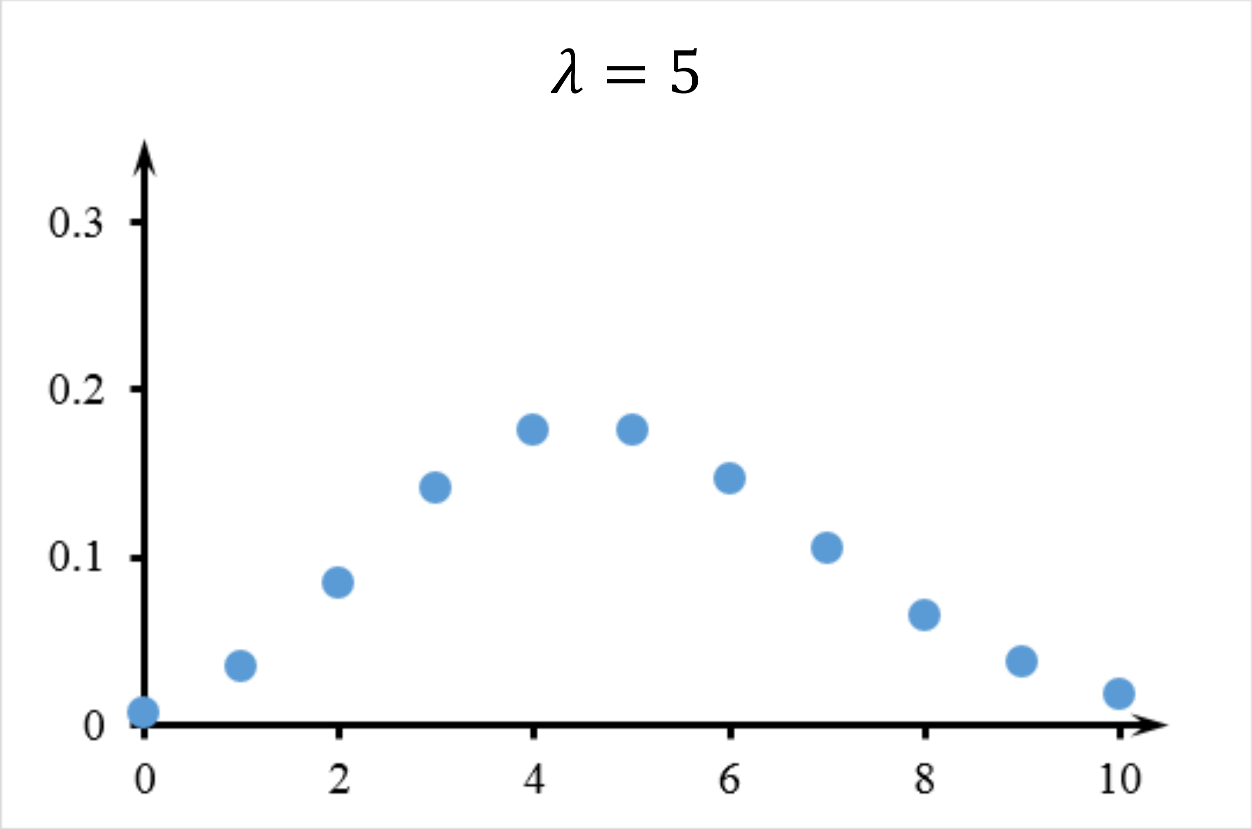
이항

푸아송과 이항



이항

푸아송과 이항



푸아송

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

1). 앞으로 1년간 한번도 shutdown 급의 기술 문제가 없을 확률은?

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

1). 앞으로 1년간 한번도 shutdown 급의 기술 문제가 없을 확률은?

단위시간을 1년으로 잡으면 $\lambda = \frac{2}{5} = 0.4$가 된다. 그러므로

$$P(0) = \frac{0.4^0 e^{-0.4}}{0!} = e^{-0.4} = 0.6703 \qquad \Rightarrow \text{대략 } 67\%$$

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

2). 앞으로 1년간 한번 이상 shutdown 급의 기술 문제가 발생할 확률은?

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

2). 앞으로 1년간 한번 이상 shutdown 급의 기술 문제가 발생할 확률은?

$P(1) + P(2) + P(3) + \cdots$?? 이 방법은 계산하기 어렵다.

그러므로, 여사건의 확률 $P(0)$을 사용해서 계산하면,

$1 - P(0) \ = \ 1 - 0.6703 \ = \ 0.3297$ $\Rightarrow$ 대략 33%

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

3). 앞으로 3년간 한번도 shutdown 급의 기술 문제가 없을 확률은?

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

3). 앞으로 3년간 한번도 shutdown 급의 기술 문제가 없을 확률은?

λ는 단위시간 안에 발생하는 사건의 평균 횟수임을 상기해 본다. 그러므로, 1년단위의 $\lambda = 0.4$이었고 단위시간이 3년으로 증가하면 $\lambda = 0.4 \times 3 = 1.2$이 되겠다.

$$P(0) = \frac{1.2^0 e^{-1.2}}{0!} = e^{-1.2} = 0.301 \quad \Rightarrow \text{ 대략 } 30\%$$

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

4). 앞으로 3년간 1~2회 shutdown 급의 기술 문제가 발생할 확률은?

푸아송 확률분포

문제: 어느 원자력 발전소에서 최근 5년간 2회 shutdown 급의 기술 문제가 있었다. 다음 물음에 답하시오.

4). 앞으로 3년간 1~2회 shutdown 급의 기술 문제가 발생할 확률은?

계속해서 $\lambda = 1.2$를 사용한다 (단위시간=3년). 구하고자 하는 확률은 다음과 같다.

$$P(1 \leq X \leq 2) = P(1) + P(2) = \frac{1.2^1 e^{-1.2}}{1!} + \frac{1.2^2 e^{-1.2}}{2!} = 0.578 \quad \Rightarrow \text{ 대략 } 57.8\%$$

연속확률 – Part 1

키포인트

- 연속확률변수와 연속확률밀도.
- 연속균등분포.

연속확률변수

- 연속확률변수 (continuous random variable): 셀 수 없는 (무한대) 가지수의 값을 가지는 확률변수.

  예). 1년 연봉, 성인남성의 신장, 등.

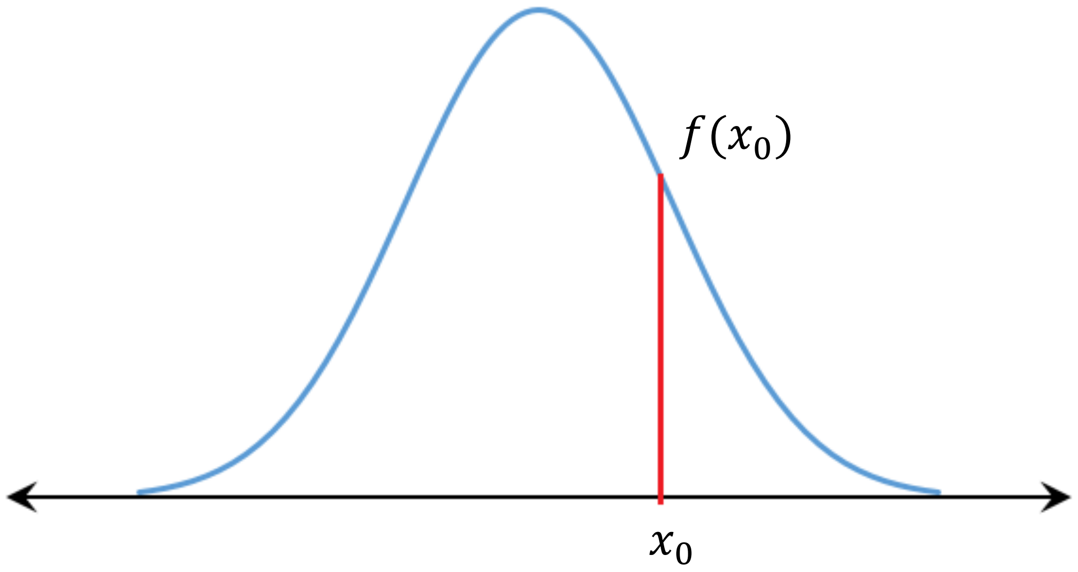
- 연속확률변수의 경우 확률은 구간에 대해서 정의되어 있다. 즉 $P(X = \boldsymbol{x_0})$와 같이 특정 위치에 대한 확률은 의미가 없고, $P(x_1 \leq X \leq x_2)$와 같이 X가 어느 구간에 있을 확률이 의미가 있다.

연속확률밀도

- 이산확률분포와는 다르게 연속확률밀도(probability density) 자체만으로는 확률의 의미가 없다.
- 이것을 사용하여 연속확률변수의 값이 특정 구간에 속할 확률을 나타낼 수 있다.
- 연속확률밀도 또는 연속확률분포를 $f(x)$와 같이 표기하여 구간의 확률 $P(x)$와는 구분 짓도록 한다. (*)

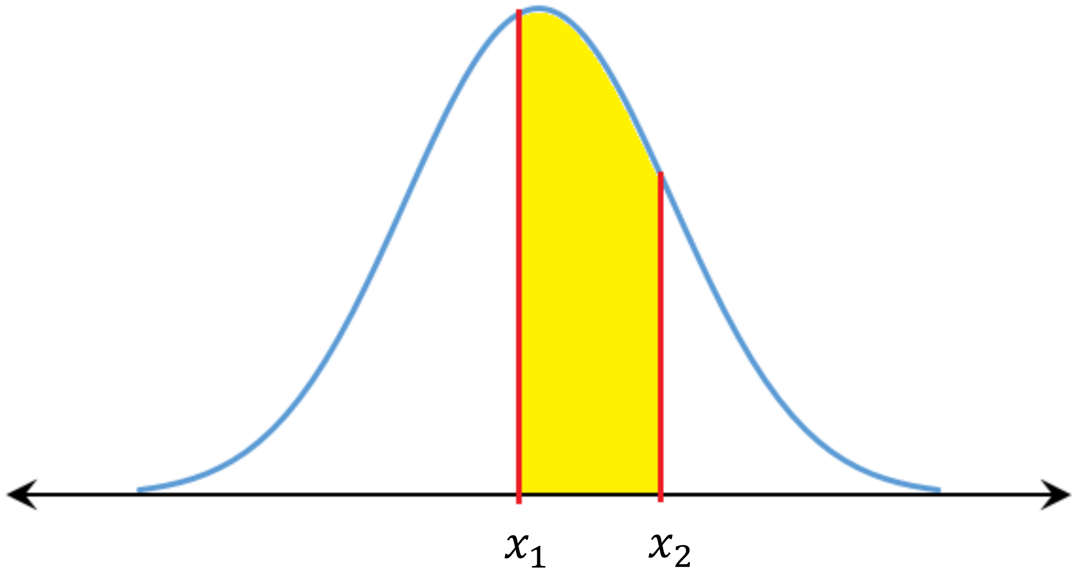
(*) "연속확률밀도" 또는 "연속확률분포" 용어를 같은 의미로 사용하도록 한다.

연속확률밀도



- $f(x_0)$에는 확률의 의미가 없다.

연속확률밀도



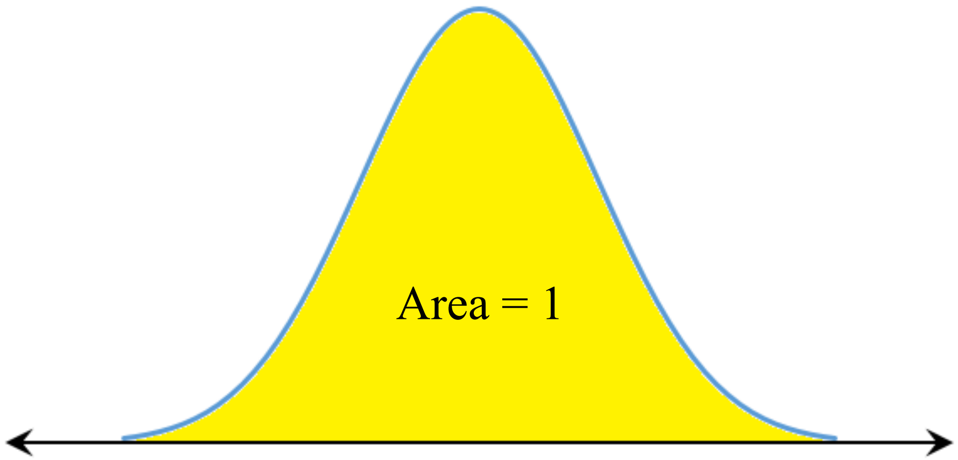
- $P(x_1 \leq X \leq x_2)$와 같이 X가 어느 구간에 있을 확률이 의미가 있다.

연속확률밀도의 필수 조건

- 다음 조건이 충족되어야 한다.

$0 \leq f(x)$

$f(x)$가 정의되어 있는 구간에서 $f(x)$ 아래의 총 면적은 1과 같아야 한다.

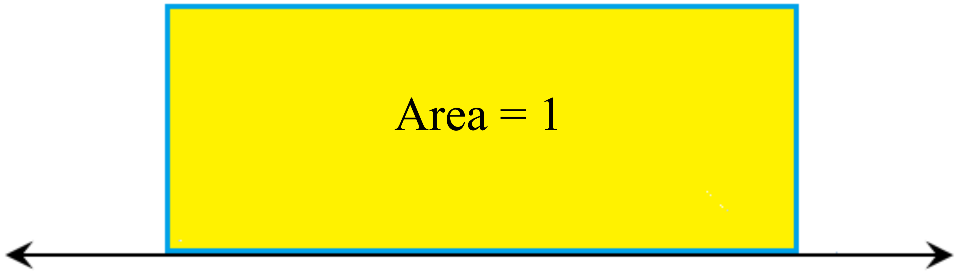


연속확률밀도의 필수 조건

- 다음 조건이 충족되어야 한다.

$0 \leq f(x)$

$f(x)$가 정의되어 있는 구간에서 $f(x)$ 아래의 총 면적은 1과 같아야 한다.



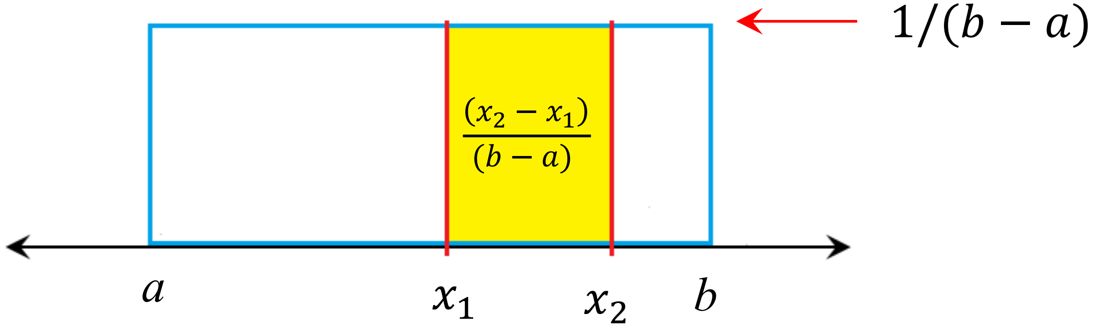
연속균등분포

- 연속균등분포 (uniform distribution)은 구간 $[a, b]$에 대해서 정의되어 있다:

$$f(x) = \frac{1}{(b-a)}$$

$\Rightarrow$ 이외의 구간에서는 $f(x) = 0$이다.

- 확률밀도가 균등하므로 확률은 구간 $[x_1, x_2]$의 폭에 비례한다.



연속균등분포

- "확률변수 X가 연속균등분포를 따른다." $\Leftrightarrow$ $X \sim Unif(a, b)$

- 평균 $= E[X] = \int_a^b \frac{x}{(b-a)} dx = \frac{1}{2}(a+b)$

- 분산 $= E[X^2] - (E[X])^2 = \int_a^b \frac{x^2}{(b-a)} dx - \left[\frac{1}{2}(a+b)\right]^2$

$$= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{1}{12}(b-a)^2$$

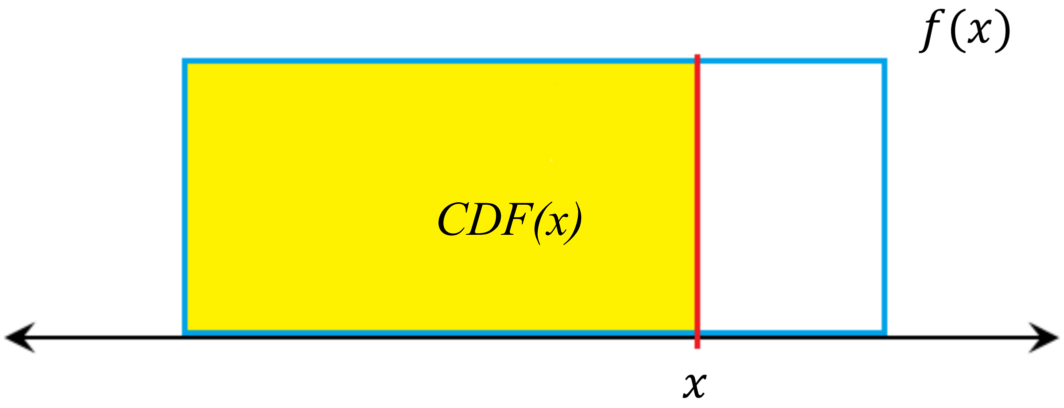
- 표준편차 $= \frac{1}{\sqrt{12}}(b-a)$

연속균등분포의 누적확률

- 연속균등분포의 누적확률 $CDF(x)$는 구간 $(-\infty, x]$ 에서 $f(x)$ 아래의 면적과 같다. 즉,

  $CDF(x) = P(-\infty < X \leq x)$이다.

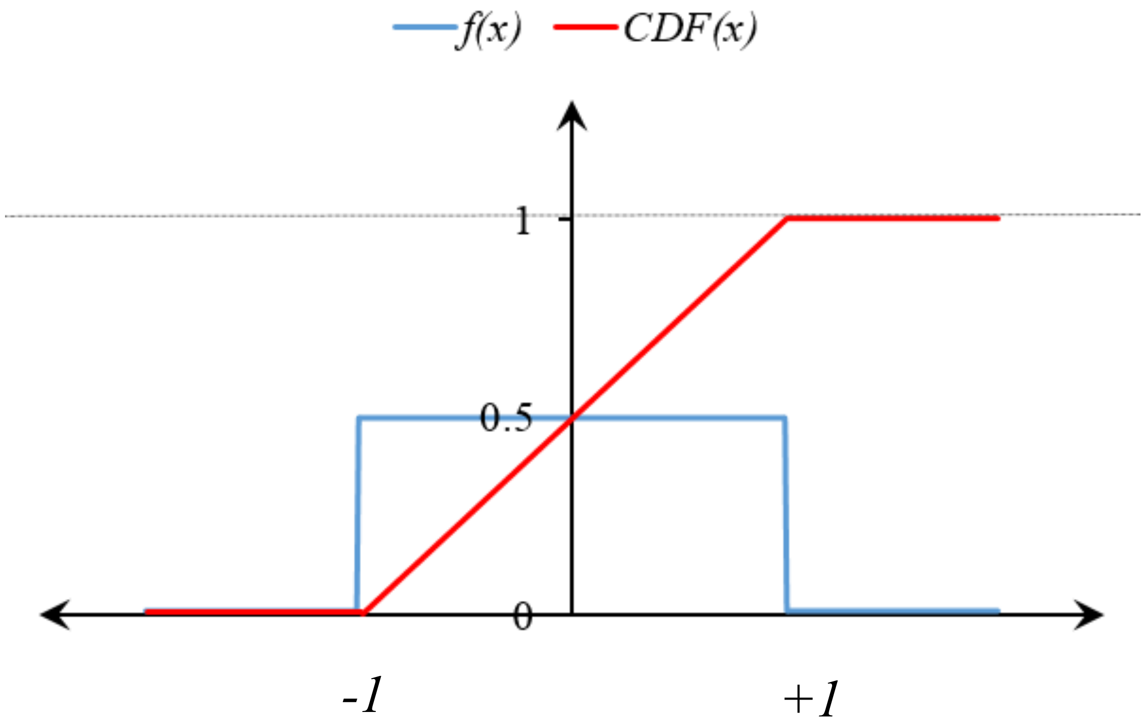
$$CDF(x) = \int_{-\infty}^{x} f(y)dy$$



연속균등분포의 누적확률

- $CDF(x)$ 는 x가 증가하면 1로 수렴한다.

예). $a = -1, b = +1$



연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

1). 백열전구가 사용시간 6000시간과 7000시간 사이에서 타버릴 확률은?

연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

1). 백열전구가 사용시간 6000시간과 7000시간 사이에서 타버릴 확률은?

확률밀도 함수는 $f(x) = \frac{1}{(7000-5000)} = 0.0005$이다. 그러므로,

$$P(6000 \leq X \leq 7000) = (7000 - 6000) \times f(x) = 1000 \times 0.0005 = 0.5$$

연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

2). 백열전구의 수명이 5500시간 이하일 확률은?

연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

2). 백열전구의 수명이 5500시간 이하일 확률은?

$$P(X \leq 5500) = P(5000 \leq X \leq 5500)$$

$$= (5500 - 5000)\times f(x) = 500\times 0.0005 = 0.25$$

연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

3). 백열전구의 수명이 최소 사용시간 5500시간 이상일 확률은?

연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

3). 백열전구의 수명이 최소 사용시간 5500시간 이상일 확률은?

$$P(5500 \le X) = P(5500 \le X \le 7000)$$

$$= (7000 - 5500) \times f(x) = 1500 \times 0.0005 = 0.75$$

연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

4). 백열전구의 수명이 정확하게 6000시간일 확률은?

연속균등분포

문제: 백열전구의 수명 X는 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

4). 백열전구의 수명이 정확하게 6000시간일 확률은?

$$P(X = 6000) = P(6000 \le X \le 6000)$$

$$= (6000 - 6000) \times f(x) = 0 \times 0.0005 = 0$$

연속확률 – Part 2

키포인트

- 정규분포.
- 정규확률변수의 합.
- 누적확률.
- 표준화와 표준정규분포.

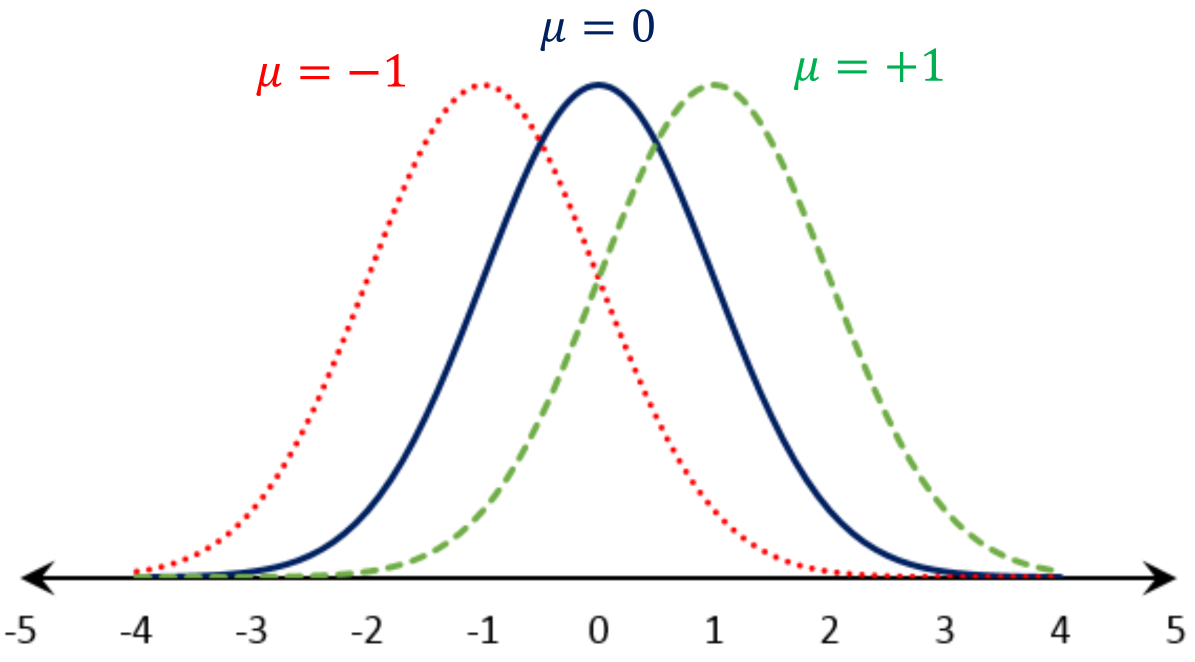
정규분포

- 정규분포 (normal distribution)은 구간 $(-\infty, +\infty)$에 대해서 정의되어 있다:

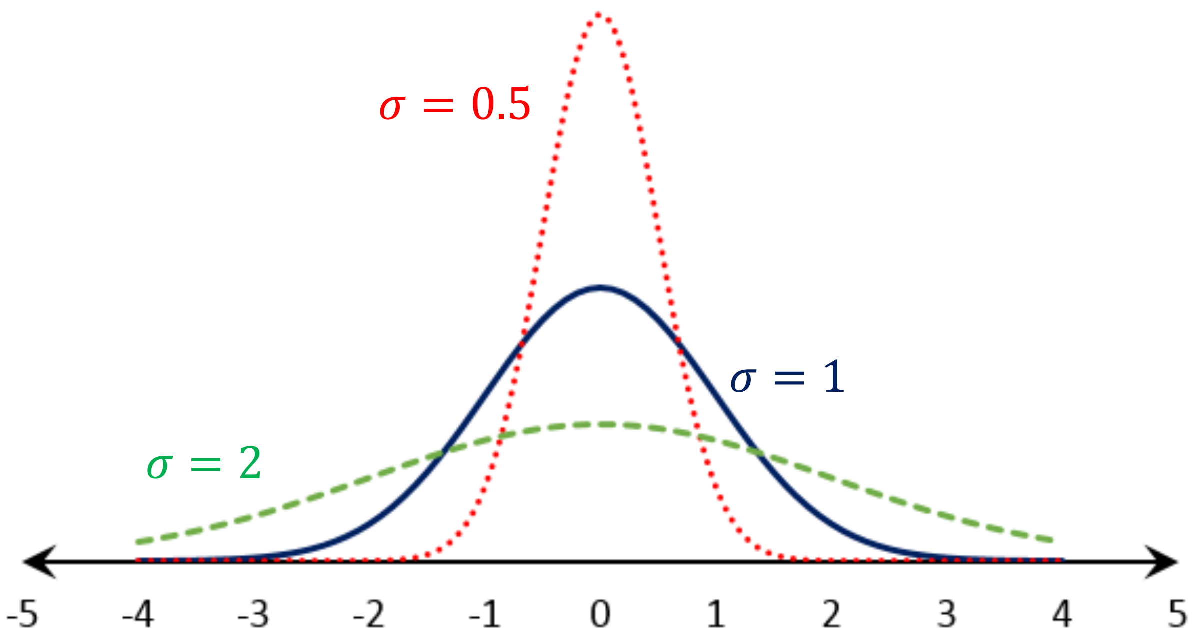
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 평균 = μ
- 분산 $= \sigma^2$
- 표준편차 = σ
- "확률변수 X가 정규확률분포를 따른다." ⇔ $X \sim N(\mu, \sigma^2)$

정규분포: μ의 역할



정규분포: σ의 역할



정규확률변수의 합

- 확률변수 X와 Y가 서로 독립이며 정규확률분포를 따를 때,

  $X \sim N(\mu_1, \sigma_1^2)$

  $Y \sim N(\mu_2, \sigma_2^2)$

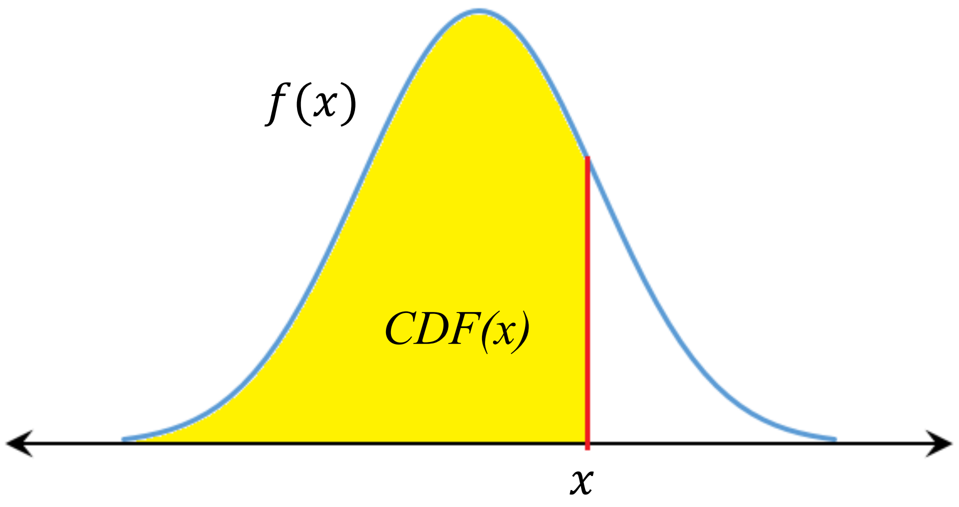
  $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$이고,

  $X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$이다.

정규분포의 누적확률

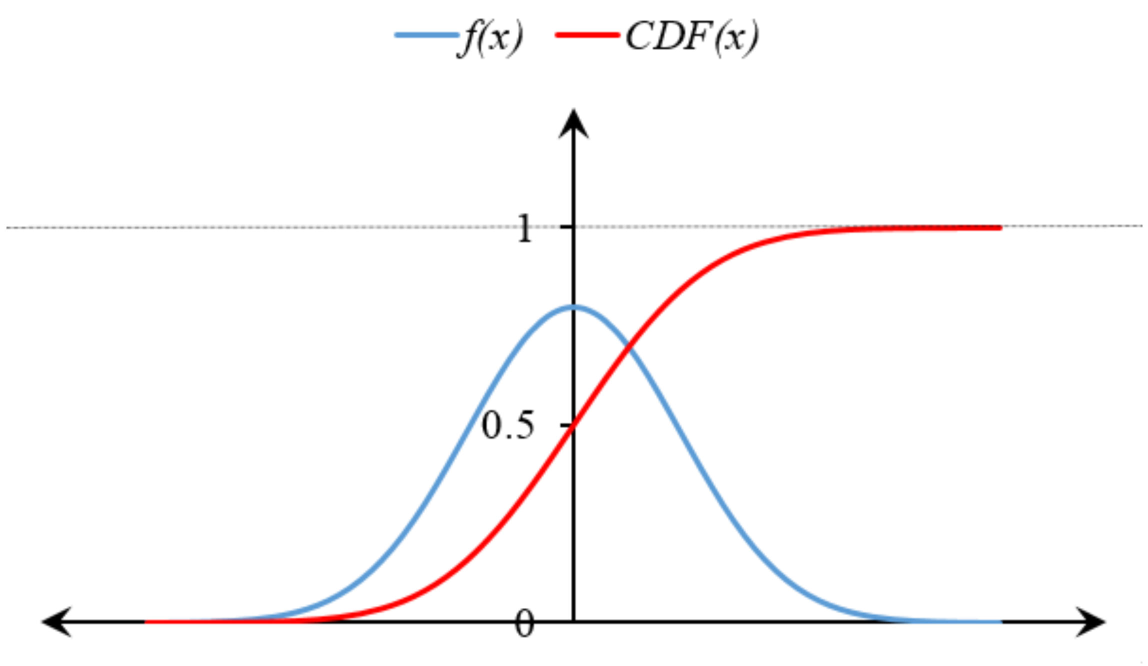
- 정규분포의 누적확률 $CDF(x)$는 구간 $(-\infty, x]$ 에서 $f(x)$ 아래의 면적과 같다. 즉, $CDF(x) = P(-\infty < X \leq x)$이다.

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



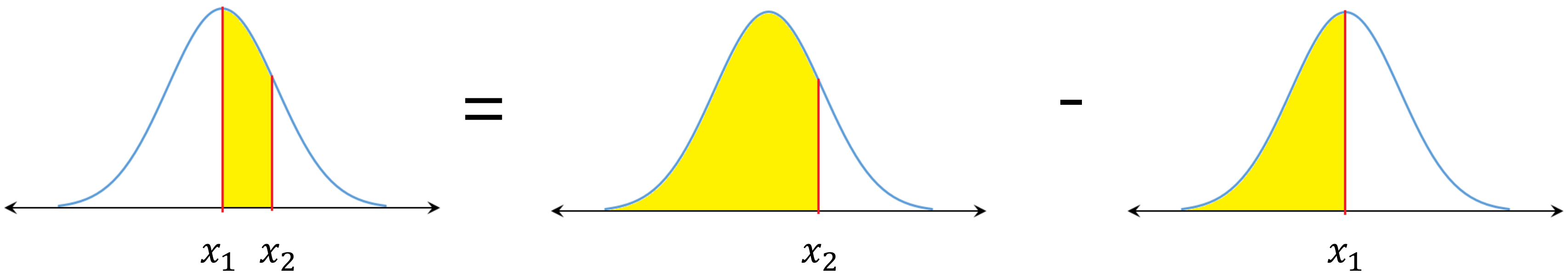
정규분포의 누적확률

- $CDF(x)$ 는 x가 증가하면 1로 수렴한다.



누적확률과 구간의 확률

- $P(x_1 \le X \le x_2)$와 같은 구간의 확률은 $CDF(x)$를 사용해서 계산할 수 있다



$$P(x_1 \le X \le x_2) \quad = \quad CDF(x_2) \quad - \quad CDF(x_1)$$

표준정규분포

- 표준정규분포는 구간 $(-\infty, +\infty)$에 대해서 정의되어 있다:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- 평균 $= 0$
- 분산 $= 1$
- 표준편차 $= 1$
- "확률변수 X가 표준정규분포를 따른다." $\Leftrightarrow$ $X \sim N(0,1)$

표준화

- 확률변수 X가 정규분포를 따르는 경우 $X \sim N(\mu, \sigma^2)$, 다음의 방식으로 X를 표준정규 확률변수로 변환할 수 있다. 그러면 $Z \sim N(0,1)$이다. 이것을 “표준화”라고 부른다.

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

- 표준화된 X의 값 z를 *z-score* 즉 “표준점수”라고 부른다.

$$z - score = \frac{x - \mu}{\sigma}$$

표준화

- 반대로 표준정규 확률변수 Z를 정규분포 $N(\mu, \sigma^2)$를 따르는 확률변수로 변환할 수도 있다.

$$X = \sigma Z + \mu$$

표준정규분포

문제: A군이 시험문제를 푸는데 문항당 평균 50초가 걸리고 표준편차는 20초라고 한다.

48초과 54초 사이에 문항을 풀 확률은 얼마인가? (다음 표준정규분포의 CDF 표 활용)

z	CDF(z)
-0.2	0.4207
-0.1	0.4602
0	**0.5**
0.1	0.5398
0.2	0.5793

표준정규분포

문제: A군이 시험문제를 푸는데 문항당 평균 50초가 걸리고 표준편차는 20초라고 한다.

48초과 54초 사이에 문항을 풀 확률은 얼마인가? (다음 표준정규분포의 CDF 표 활용)

z	CDF(z)
-0.2	0.4207
-0.1	0.4602
0	**0.5**
0.1	0.5398
0.2	0.5793

먼저 $x_1 = 48$초와 $x_2 = 54$초를 표준화 한다.

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{48-50}{20} = -\frac{2}{20} = -0.1 \quad , \quad z_2 = \frac{x_2 - \mu}{\sigma} = \frac{54-50}{20} = \frac{4}{20} = 0.2$$

표준정규분포

문제: A군이 시험문제를 푸는데 문항당 평균 50초가 걸리고 표준편차는 20초라고 한다.

48초과 54초 사이에 문항을 풀 확률은 얼마인가? (다음 표준정규분포의 CDF 표 활용)

z	CDF(z)
-0.2	0.4207
-0.1	0.4602
0	**0.5**
0.1	0.5398
0.2	0.5793

그리고, CDF 표를 활용하여 확률을 계산한다.

$$P(z_1 \leq Z \leq z_2) = CDF(z_2) - CDF(z_1) = CDF(0.2) - CDF(-0.1)$$
$$= 0.5793 - 0.4602 = 0.1191$$

연속확률 - Part 3

키포인트

- 지수확률분포.
- 카이제곱 확률분포.
- 카이제곱 확률변수의 합.

지수확률분포

- 지수확률분포 (exponential distribution)은 푸아송 사건 사이의 거리(시간)을 확률로 모델링하는 함수이다.

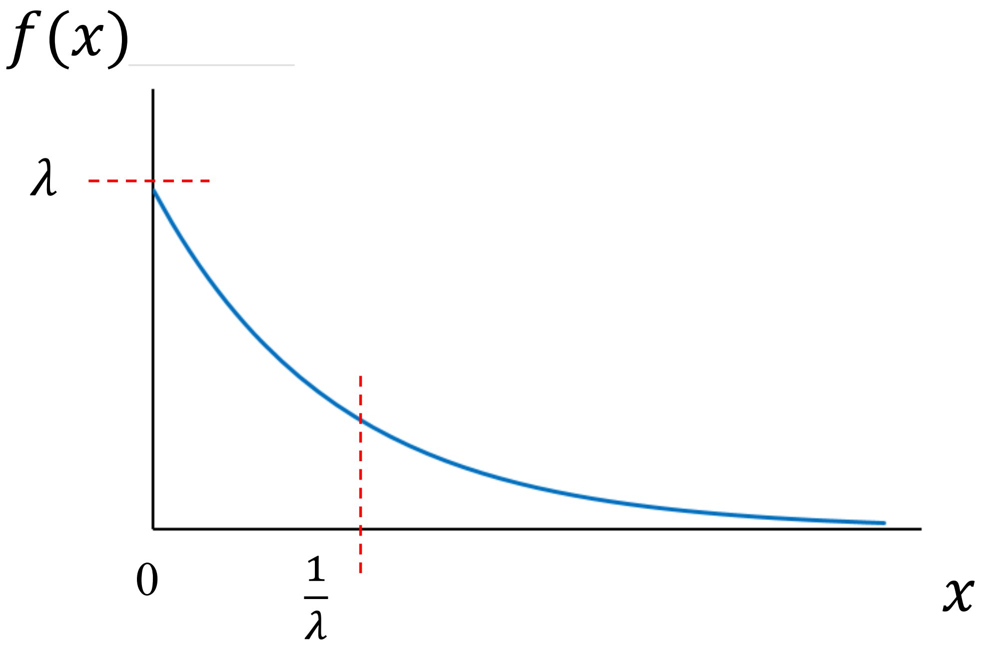
$$f(x) = \lambda e^{-\lambda x}$$

- x는 양의 실수이어야 한다 ($x \geq 0$).
- λ는 유일한 파라미터이고 양의 수치여야 한다.
- 보통 $\frac{1}{\lambda}$는 시간의 의미를 갖는다. $\Rightarrow$ 푸아송 사건과 사건 사이의 시간적 거리.
- "확률변수 X가 지수확률분포를 따른다." $\Leftrightarrow$ $X \sim Exp(\lambda)$

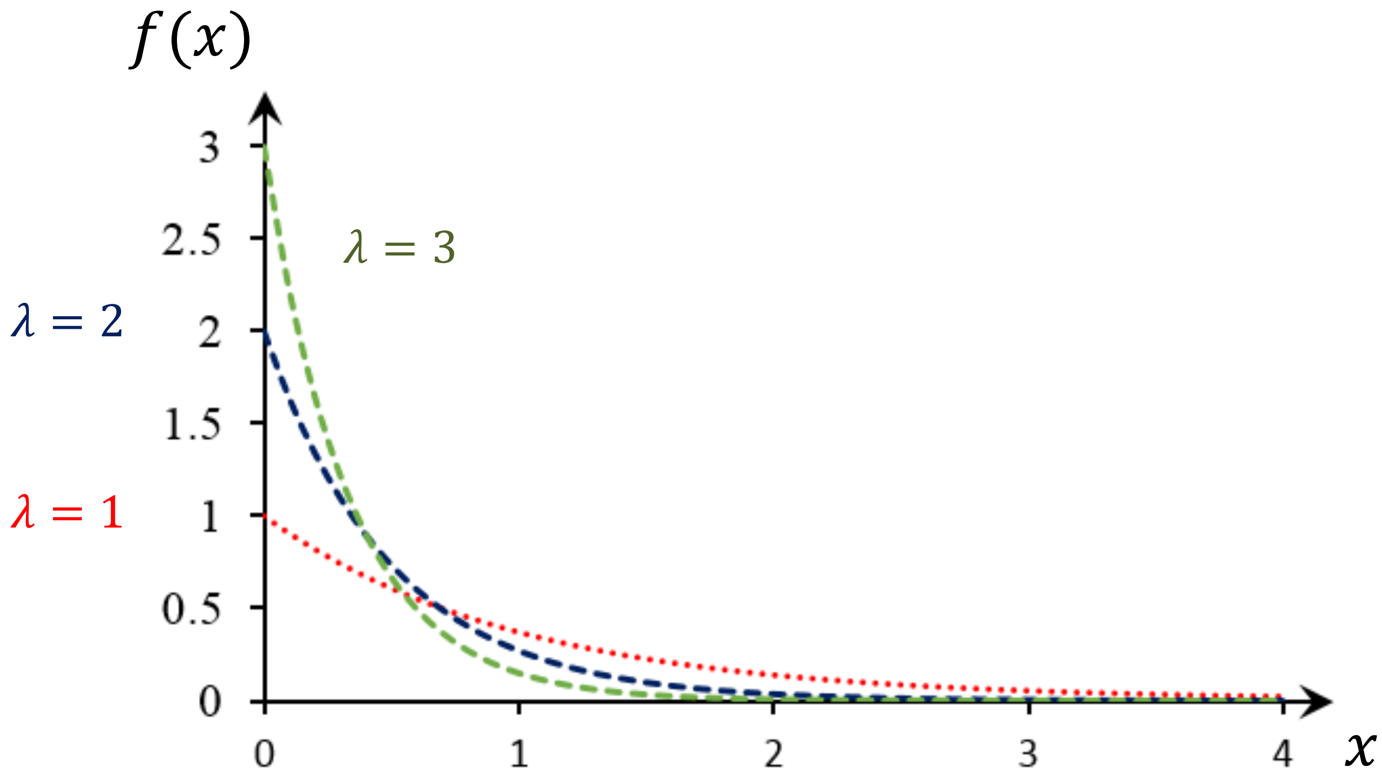
지수확률분포

- 평균 $= \frac{1}{\lambda}$
- 분산 $= \frac{1}{\lambda^2}$
- 표준편차 $= \frac{1}{\lambda}$
- 지수분포의 누적확률: $CDF(x) = 1 - e^{-\lambda x}$

지수확률분포: λ의 역할



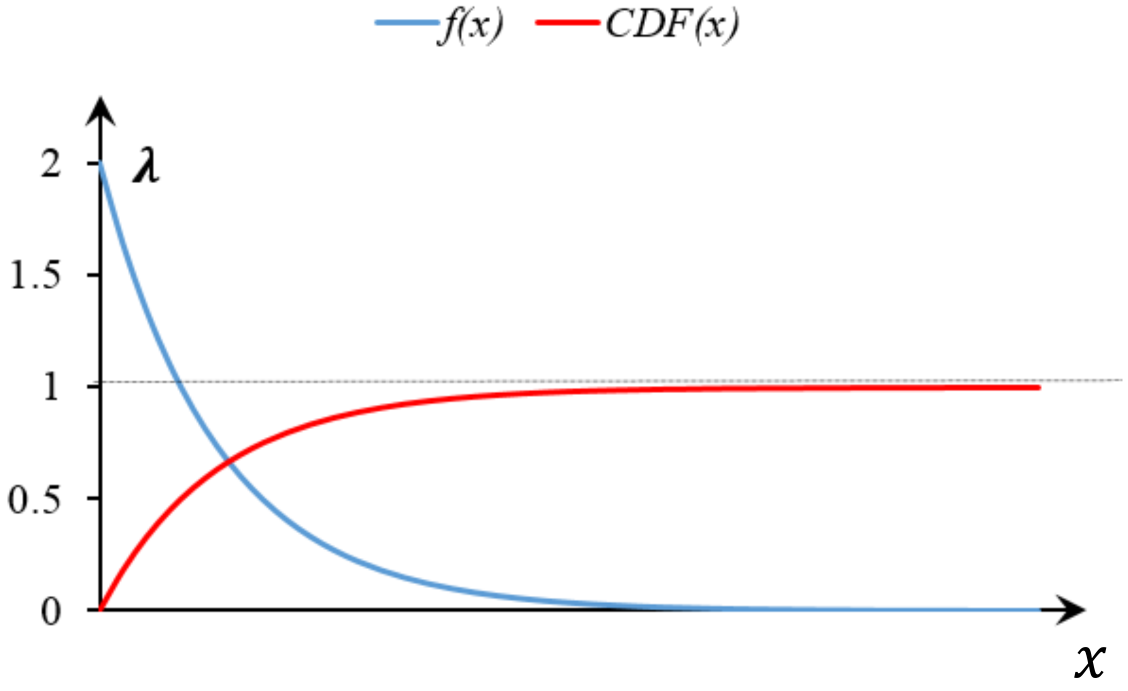
지수확률분포: λ의 역할



지수확률분포의 누적확률

- $CDF(x)$는 x가 증가하면 1로 수렴한다.

예). $\lambda = 2$



지수확률분포

문제: 자동차가 평균 20개월 마다 한번씩 고장 난다는 전제로 지수확률분포로 모델링하여 고장 주기가 15개월과 20개월 사이일 확률을 구하시오.

지수확률분포

문제: 자동차가 평균 20개월 마다 한번씩 고장 난다는 전제로 지수확률분포로 모델링하여 고장 주기가 15개월과 20개월 사이일 확률을 구하시오.

먼저 $\lambda = \frac{1}{20} = 0.05$, 그러므로 CDF를 활용하여 확률을 계산한다.

$$P(15 \leq X \leq 20) = CDF(20) - CDF(15) = (1 - e^{-0.05\times 20}) - (1 - e^{-0.05\times 15})$$

$$= 0.6321 - 0.5276 = 0.1045$$

카이제곱 확률분포

- k개의 표준정규분포를 따르는 독립적인 확률변수 $Z_i \sim N(0,1)$가 있을때 카이제곱 확률변수 (chi-square) Q는 이들의 제곱의 합으로 정의한다.

$$Q = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

- 여기에서 k를 "자유도"라고 부른다.
- "확률변수 Q가 카이제곱 확률분포를 따른다." $\Leftrightarrow$ $Q \sim \chi^2(k)$

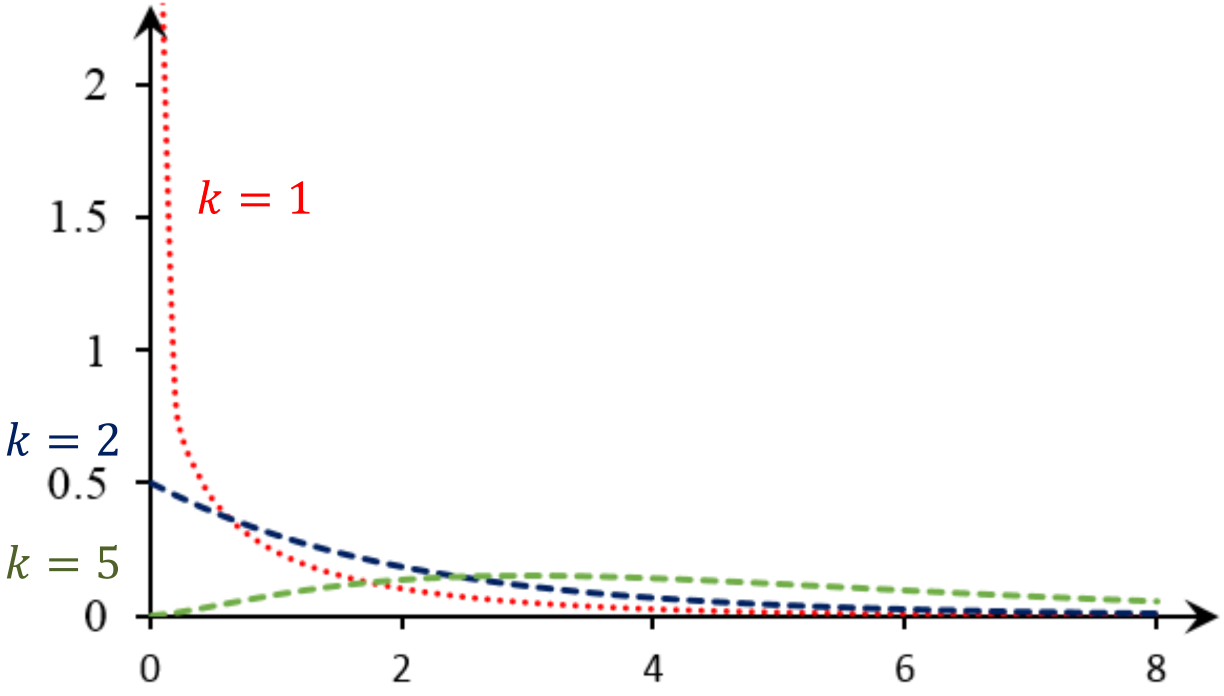
카이제곱 확률분포

- 카이제곱 확률분포함수는 구간 $(0, +\infty)$에 대해서 정의되어 있다:

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

- 평균 $= k$
- 분산 $= 2k$
- 표준편차 $= \sqrt{2k}$

카이제곱 확률분포: 자유도 k의 역할



카이제곱 확률변수의 합

- 확률변수 Q_1와 Q_2가 서로 독립적이며 다음과 같이 카이제곱 확률분포를 따를 때,

$$Q_1 \sim \chi^2(\boldsymbol{k_1})$$

$$Q_2 \sim \chi^2(\boldsymbol{k_2})$$

$Q_1 + Q_2 \sim \chi^2(\boldsymbol{k_1} + \boldsymbol{k_2})$이다.

카이제곱 확률변수의 합

- 이전 페이지의 증명은 다음과 같이 매우 간단하다.

$$Q_1 + Q_2 = \{Z^2 + Z^2 + \cdots + Z^2\} + \{Z^2 + \cdots + Z^2\}$$

$\leftarrow$ $\boldsymbol{k_1}$ 개 $\rightarrow$ $\leftarrow$ $\boldsymbol{k_2}$ 개 $\rightarrow$

그러므로, $Q_1 + Q_2 \sim \chi^2(\boldsymbol{k_1} + \boldsymbol{k_2})$.

연속확률 – Part 4

키포인트

- 스튜던트 t 확률분포.
- F 확률분포.

스튜던트 t 확률분포

- $Q \sim \chi^2(\nu)$이고 $Z \sim N(0,1)$일때 스튜던트 t 확률변수 T는 다음과 같이 정의 된다.

$$T = \frac{Z}{\sqrt{Q/\nu}}$$

- 여기에서 ν는 카이제곱 확률변수의 "자유도"이다.
- "확률변수 T가 스튜던트 t 확률분포를 따른다." $\Leftrightarrow$ $T \sim t(\nu)$
- 자유도 ν가 커질수록 스튜던트 t는 표준정규분포로 수렴한다.

스튜던트 t 확률분포

- 스튜던트 t 확률분포는 구간 $(-\infty, +\infty)$에 대해서 정의되어 있다:

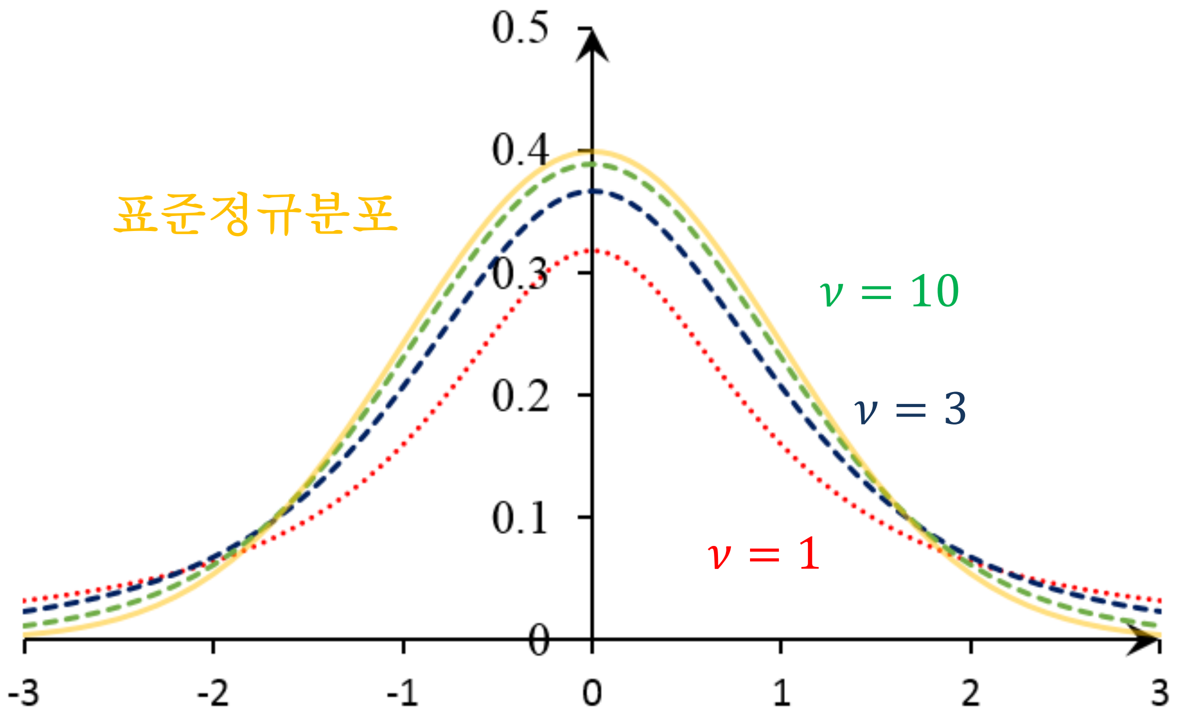
$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- 평균 = *0*
- 분산 $= \frac{\nu}{\nu-2}$ $\Leftarrow$ $\nu > 2$인 경우에만.
- 표준편차 $= \sqrt{\frac{\nu}{\nu-2}}$ $\Leftarrow$ $\nu > 2$인 경우에만.

스튜던트 t 확률분포

- 스튜던트 t 분포함수는 통계에서 신뢰구간을 계산하는데 매우 중요한 역할을 한다.
- 표본에서 추출한 분산이 모분산과 크게 다를때 (표본의 크기가 작을 때), 표준정규분포의 분위수 대신에 스튜던트 t 분포의 분위수를 사용해서 신뢰구간을 계산한다.
- 표본의 크기가 n일때 자유도는 $\nu = n - 1$이다.
- 표본의 크기가 커진다는 것은 자유도가 증가한다는 것과 같은 의미이고, 이때 스튜던트 t는 표준정규분포에 수렴하게 되니 표준정규분포의 분위수나 스튜던트 t의 분위수 큰 차이가 없게 된다.

스튜던트 t 확률분포: 자유도 ν의 역할



F 확률분포

- $Q_1 \sim \chi^2(d_1)$이고 $Q_2 \sim \chi^2(d_2)$일때 F 확률변수 X는 다음과 같이 정의 된다.

$$X = \frac{Q_1/d_1}{Q_2/d_2}$$

- 여기에서 d_1와 d_2는 카이제곱 확률변수의 “자유도”이다:

$\Rightarrow\ d_1 =$ 분자의 자유도

$\Rightarrow\ d_2 =$ 분모의 자유도

F 확률분포

- "확률변수 X가 F 확률분포를 따른다". $\Leftrightarrow$ $X \sim F(d_1, d_2)$
- F 검정, 분산분석 (ANOVA) 등에 활용된다.
- 자유도 d_2가 커질수록 F분포는 카이제곱 ${}^{Q_1}/_{d_1}$로 수렴한다.

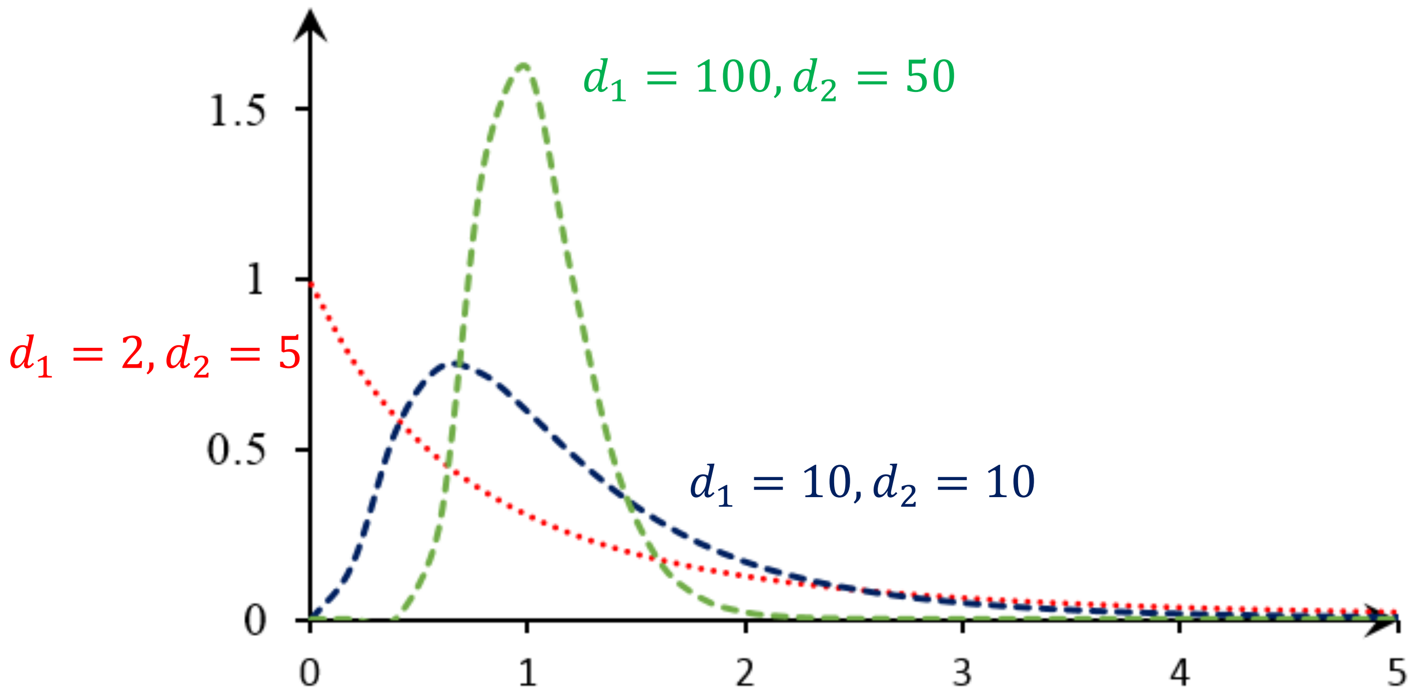
F 확률분포

- 평균 $= \dfrac{d_2}{d_2-2}$ $\Leftarrow$ $d_2 > 2$인 경우

- 분산 $= \dfrac{2d_2^2(d_1+d_2-2)}{d_1(d_2-4)(d_2-2)^2}$ $\Leftarrow$ $d_2 > 4$인 경우

- 최빈값 (mode) $= \left(\dfrac{d_1-2}{d_1}\right)\left(\dfrac{d_2}{d_2+2}\right)$ $\Leftarrow$ $d_1 > 2$인 경우

F 확률분포



결합확률

키포인트

- 결합확률.
- 이변량 결합확률분포.
- 공분산과 상관계수.
- 독립성과 상관성.
- 변수의 합.

이변량 결합확률분포

- 두개의 확률변수 X와 Y 에 대한 확률. (*)
- 이변량 결합확률분포: $P(x, y) = P(X = x, Y = y)$.

$\Rightarrow X = x$ **AND** $Y = y$ 일 확률.

(*) 여기에서는 이산확률의 경우를 전제하는데, 연속확률의 경우에도 유사한 방법으로 이변량 결합확률분포를 정의할 수 있다.

공분산

- $Cov(X, Y) = E\big[(X - \mu_x)\big(Y - \mu_y\big)\big]$

$$= \sum_{all\ x_i} \sum_{all\ y_i} (x_i - \mu_x)(y_i - \mu_y) P(x_i, y_i)$$

- 공분산의 간편 수식: $Cov(X, Y) = E[X\,Y] - E[X]\,E[Y]$

$$= \sum_{all\ x_i} \sum_{all\ y_i} x_i\, y_i P(x_i, y_i) - \mu_x \mu_y$$

- $Var(X) = Cov(X, X)$ $\Leftarrow$ 분산과 공분산의 연결.

상관계수

- $Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$
- 상관계수의 값은 -1과 1사이의 수치이다.
- 상관계수는 선형관계의 방향과 강도를 나타낸다.

$\Rightarrow Cor(X,Y) > 0$: X와 Y사이에 양의 선형관계가 있음.

$\Rightarrow Cor(X,Y) < 0$: X와 Y사이에 음의 선형관계가 있음.

$\Rightarrow Cor(X,Y) = 0$: X와 Y사이에 선형관계가 **없음**.

독립성과 상관성

- 독립성: $P(X, Y) = P(X)\, P(Y)$.

  $\Rightarrow Cov(X, Y) = E[X\, Y] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0$.

  $Cor(X, Y) = 0$. 그러므로 “상관성 없음”을 내포한다.

- 상관계수: $Cor(X, Y)$.

  ⇒ 상관계수는 -1과 1 사이의 수치이다.

  ⇒ “상관성이 없다” = “상관계수 0”의 의미. 하지만 독립성을 내포하지는 않는다.

독립성과 상관성

예: X와 Y 동전 두 개를 동시에 던지는 경우를 가정해 본다. 표본공간은 HH, HT, TH, TT 로 이루어져 있다. 독립성을 확인해 본다.

개개 확률분포함수는 다음과 같다:

$$\left.\begin{aligned} P(X=H) &= \frac{1}{2} \\ P(X=T) &= \frac{1}{2} \end{aligned}\right\} \quad P(X)$$

$$\left.\begin{aligned} P(Y=H) &= \frac{1}{2} \\ P(Y=T) &= \frac{1}{2} \end{aligned}\right\} \quad P(Y)$$

독립성과 상관성

예: X와 Y 동전 두 개를 동시에 던지는 경우를 가정해 본다. 표본공간은 HH, HT, TH, TT로 이루어져 있다. 독립성을 확인해 본다.

다음과 같이 $P(X,Y) = P(X)P(Y)$을 확인할 수 있다. 즉, X와 Y는 서로 독립관계이다.

$$P(X = H, Y = H) = \frac{1}{4} = P(X = H) \times P(Y = H)$$

$$P(X = H, Y = T) = \frac{1}{4} = P(X = H) \times P(Y = T)$$

$$P(X = T, Y = H) = \frac{1}{4} = P(X = T) \times P(Y = H)$$

$$P(X = T, Y = T) = \frac{1}{4} = P(X = T) \times P(Y = T)$$

독립성과 상관성

문제: -1, 0, 1에서 동일한 확률 1/3 을 갖는 확률변수 X와 $Y = X^2$과 같이 정의되는 Y변수가 있다. 그러면 결합확률은 다음과 같다. 물음에 답하여라.

$$P(X = -1, Y = 1) = \frac{1}{3}\ ,\ P(X = 0, Y = 0) = \frac{1}{3}\ ,\ P(X = 1, Y = 1) = \frac{1}{3}$$

1). X와 Y가 서로 독립인지 확인해 보아라.

독립성과 상관성

문제: -1, 0, 1에서 동일한 확률 1/3 을 갖는 확률변수 X와 $Y = X^2$과 같이 정의되는 Y변수가 있다. 그러면 결합확률은 다음과 같다. 물음에 답하여라.

$P(X = -1, Y = 1) = \frac{1}{3}$, $P(X = 0, Y = 0) = \frac{1}{3}$, $P(X = 1, Y = 1) = \frac{1}{3}$

1). X와 Y가 서로 독립인지 확인해 보아라.

X의 확률분포함수는 다음과 같다:

$P(X = -1) = \frac{1}{3}$, $P(X = 0) = \frac{1}{3}$, $P(X = 1) = \frac{1}{3}$

그리고 Y의 확률분포함수는 다음과 같다:

$P(Y = 0) = \frac{1}{3}$, $P(Y = 1) = \frac{2}{3}$

독립성과 상관성

문제: -1, 0, 1에서 동일한 확률 1/3 을 갖는 확률변수 X와 $Y = X^2$과 같이 정의되는 Y변수가 있다. 그러면 결합확률은 다음과 같다. 물음에 답하여라.

$P(X = -1, Y = 1) = \frac{1}{3}$, $P(X = 0, Y = 0) = \frac{1}{3}$, $P(X = 1, Y = 1) = \frac{1}{3}$

1). X와 Y가 서로 독립인지 확인해 보아라.

그런데 다음을 확인할 수 있다.

$$P(X = -1, Y = 1) = \frac{1}{3} \neq P(X = -1) \times P(Y = 1) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$$

$$P(X = 0, Y = 0) = \frac{1}{3} \neq P(X = 0) \times P(Y = 0) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$$

$$P(X = 1, Y = 1) = \frac{1}{3} \neq P(X = 1) \times P(Y = 1) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$$

$\Rightarrow P(X, Y) \neq P(X)P(Y)$ 이므로 X와 Y는 서로 독립이 아닌 종속 관계이다.

독립성과 상관성

문제: -1, 0, 1에서 동일한 확률 1/3 을 갖는 확률변수 X와 $Y = X^2$과 같이 정의되는 Y변수가 있다. 그러면 결합확률은 다음과 같다. 물음에 답하여라.

$P(X = -1, Y = 1) = \frac{1}{3}$, $P(X = 0, Y = 0) = \frac{1}{3}$, $P(X = 1, Y = 1) = \frac{1}{3}$

2). 이제는 X와 Y 사이의 상관계수를 계산해 보아라.

독립성과 상관성

문제: -1, 0, 1에서 동일한 확률 1/3 을 갖는 확률변수 X와 $Y = X^2$과 같이 정의되는 Y변수가 있다. 그러면 결합확률은 다음과 같다. 물음에 답하여라.

$P(X = -1, Y = 1) = \frac{1}{3}$, $P(X = 0, Y = 0) = \frac{1}{3}$, $P(X = 1, Y = 1) = \frac{1}{3}$

2). 이제는 X와 Y 사이의 상관계수를 계산해 보아라.

$$E[X] = -1\times P(X = -1) + 0\times P(X = 0) + 1\times P(X = 1) = -\frac{1}{3} + 0 + \frac{1}{3} = 0$$

$$E[Y] = 0\times P(Y = 0) + 1\times P(Y = 1) = 0 + 1\times\frac{2}{3} = \frac{2}{3}$$

$$E[X\,Y] = E[X\,X^2] = E[X^3] = -1\times P(X = -1) + 0\times P(X = 0) + 1\times P(X = 1) = 0$$

그러므로 $Cov(X, Y) = E[X\,Y] - E[X]\,E[Y] = 0 - 0\times\frac{2}{3} = 0$

$\Rightarrow$ 공분산이 0이니 상관계수도 0이다! 서로 독립은 아닌데 상관계수가 0인 경우 발견!

두 확률변수의 합 (이변량)

- 확률변수 X와 Y의 합의 경우를 생각해 본다.
- 평균 (기대값): $E[X+Y] = E[X] + E[Y]$
- 분산: $Var(X+Y) = E\left[\left(X+Y-\mu_x-\mu_y\right)^2\right]$

$$= E\left[\left((X-\mu_x)+(Y-\mu_y)\right)^2\right]$$

$$= E\left[(X-\mu_x)^2+\left(Y-\mu_y\right)^2+2(X-\mu_x)\left(Y-\mu_y\right)\right]$$

$$= Var(X)+Var(Y)+2Cov(X,Y)$$

$= Var(X)+Var(Y)$ $\Leftarrow$ X와 Y가 서로 독립인 경우.

여러 확률변수의 합 (다변량)

- 평균 (기대값):

$$E[X_1 + X_2 + X_3 + \cdots] = E[X_1] + E[X_2] + E[X_3] + \cdots$$

- 분산:

$$Var(X_1 + X_2 + X_3 + \cdots) = Var(X_1) + Var(X_2) + Var(X_3) + \cdots$$
$$+2Cov(X_1, X_2) + 2Cov(X_1, X_3) + 2Cov(X_2, X_3) + \cdots$$

- 위에서 변수들이 서로 독립적이라면 공분산 항은 필요 없다. 그러면,

$$Var(X_1 + X_2 + X_3 + \cdots) = Var(X_1) + Var(X_2) + Var(X_3) + \cdots$$

끝