선수 지식 - 통계
연속확률분포

연속확률분포 | 딥러닝의 기초가 되는 확률 개념 알아보기

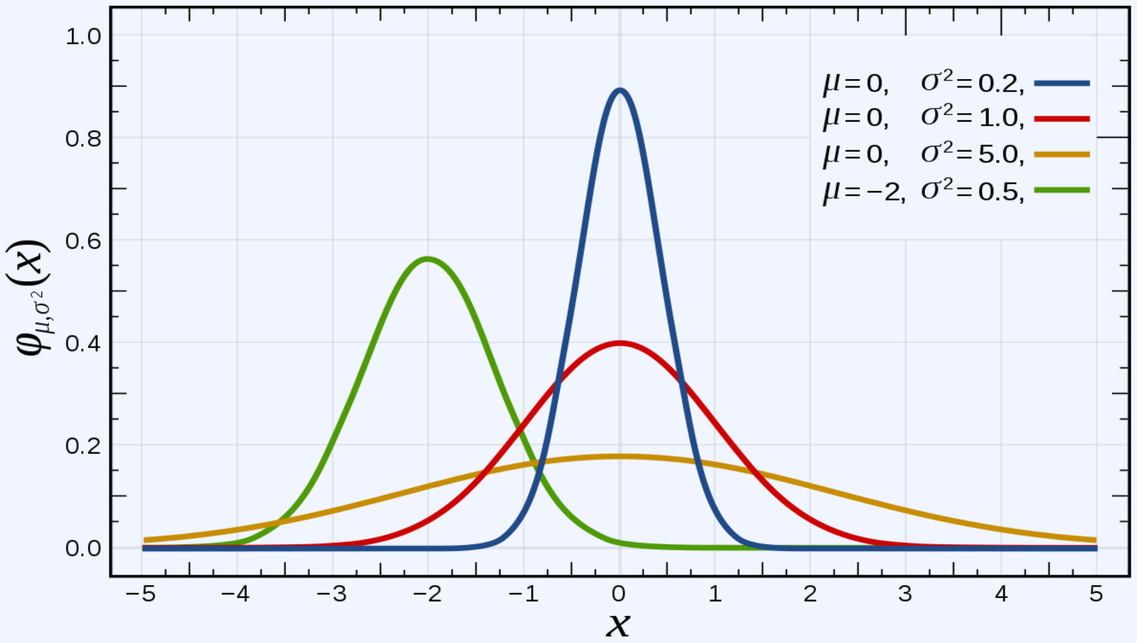
강사 나동빈

선수 지식 - 통계

연속확률분포

연속확률분포(Continuous Probability Distribution)

- 확률변수 X가 취할 수 있는 값이 무한한 경우, 이를 **연속확률변수**라고 한다.
- 이때 **연속확률분포**는 연속확률변수의 확률 분포를 의미한다.
- 대표적인 예시인 정규분포를 확인해 보자. → 매우 자주 등장하는 개념이므로, 숙지할 필요가 있다.



연속확률분포(Continuous Probability Distribution)

- 확률변수 X가 취할 수 있는 값이 무한한 경우, 이를 **연속확률변수**라고 한다.
- 연속확률분포는 셀 수 없이 많은 확률 변수들의 분포이다.

→ 특정한 값 x에 대한 정확한 확률 값을 표현할 수 없다.

→ 따라서 특정한 구간 $a \leq x \leq b$에 대한 확률로 표현한다.

확률밀도함수(Probability Density Function)

- 연속확률변수가 주어진 구간 내에 포함될 확률을 출력하는 함수다.

[참고] 이산확률변수는 확률밀도함수(PDF)가 아닌 확률질량함수(PMF)를 사용한다.

연속확률분포의 성질

1. 확률 변수 X가 어떠한 구간에 속할 확률은 0과 1사이이다.
2. 확률 변수 X가 값을 가질 수 있는 모든 구간의 확률을 합치면 1이다. (전체 면적 = 1)

→ 단, 각 구간은 배반(서로 겹치는 게 없을 때) 관계일 때 이것이 성립한다.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$ ➡ 전체 면적의 합은 1

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$ ➡ a부터 b까지의 면적

연속확률분포의 성질

- 확률밀도함수를 어떻게 이해할 수 있을까?
- $f(x)$는 **확률**을 의미하고, dx는 **구간 길이**를 의미한다.
- 다시 말해 a부터 b까지의 구간에 대하여 "확률 / 구간 길이"의 값을 모두 더한 값이다.
- 구간 길이를 **부피**로 볼 때, 전체 공식은 "질량 / 부피"이므로, 이는 **밀도**를 의미하게 된다.

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

연속확률분포 - 참고 사항

- 확률 변수 X의 값이 구간 $[a, b]$에 속할 확률은 다음과 같이 표현한다.

$$P(a \leq X \leq b)$$

- 연속 확률 변수는 면적으로 계산되며, 한 점에 대한 확률은 0으로 간주한다.
- **예시)** 나랑 키가 소수점 아래까지 **"완벽히"** 동일한 사람이 존재할 수 있을까?

$$P(a \leq X \leq b) = P(a < X < b)$$

균등 분포(Uniform Distribution)

- 가장 단순한 연속확률분포로, 특정 구간 내 값들이 나타날 가능성이 균등하다.
- 다시 말해, **모든 확률변수에 대해 일정한 확률**을 가지는 확률 분포다.
- X가 균등 분포를 따를 때 $X \sim U(a, b)$로 표현한다.
- X는 a에서 b 사이에서 일정한 값을 취하고, $P(a \leq X \leq b) = 1$이다.
- 균등 분포를 따르는 확률 변수 X의 확률밀도함수는 다음과 같다.

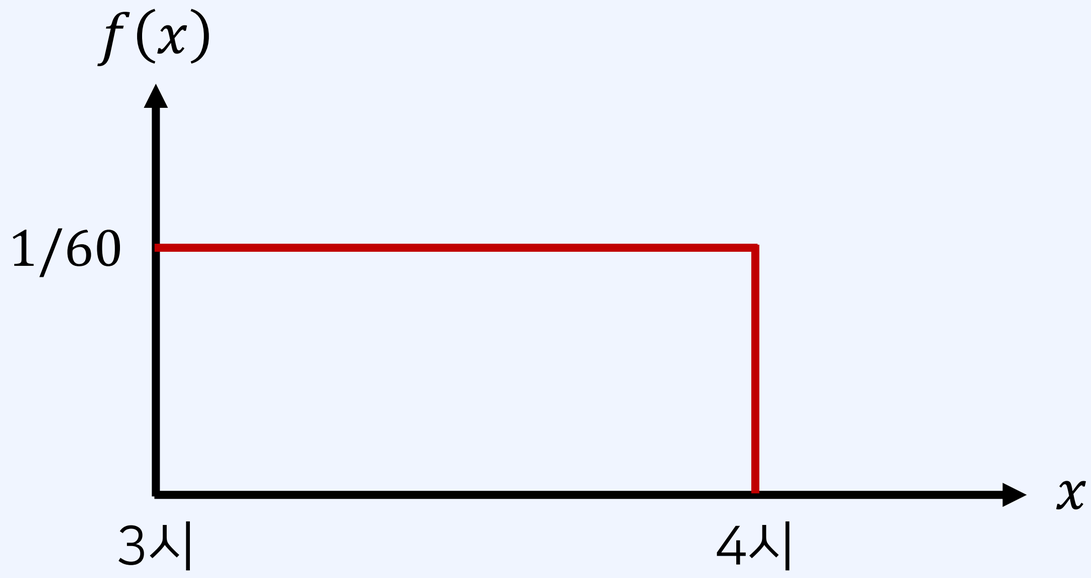
$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq X \leq b \\ 0, & otherwise \end{cases}$$

균등 분포 예시 – 스팸 메일

- 매일 오후 3시부터 오후 4시 사이에 하나의 **스팸 메일**이 도착한다.
- 해당 시간 안에서 특정 시간에 스팸 메일이 도착할 확률이 동일하다고 해보자.
- 다시 말해 오후 3시부터 오후 4시 사이에서 균등한 확률 값을 가진다.
- 이때 스팸 메일이 도착하는 시각을 확률변수 X라고 해보자.
- **확률 분포 함수**는 어떻게 그릴 수 있을까?

균등 분포 예시 – 스팸 메일

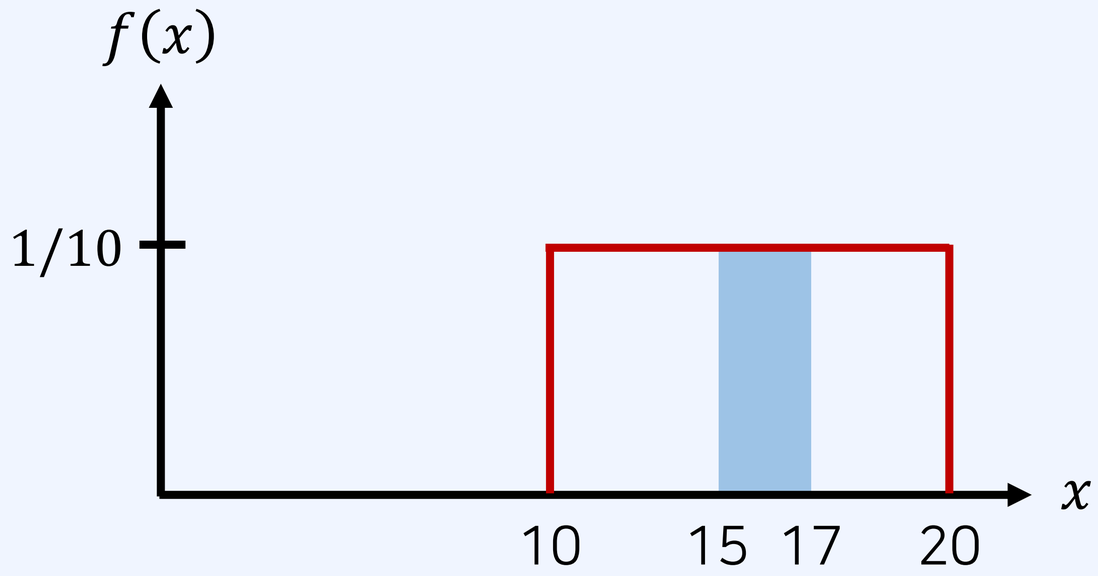
- **균등 분포**는 직사각형 형태를 보인다.
- 가로 축의 단위를 **"분"**으로 했을 때, 가로 길이가 60이 된다.
- 이때 오후 3시부터 오후 4시 시간대의 각 분에 대하여 $f(x)$는 $1/60$이다.



균등 분포 예시 – 스팸 메일

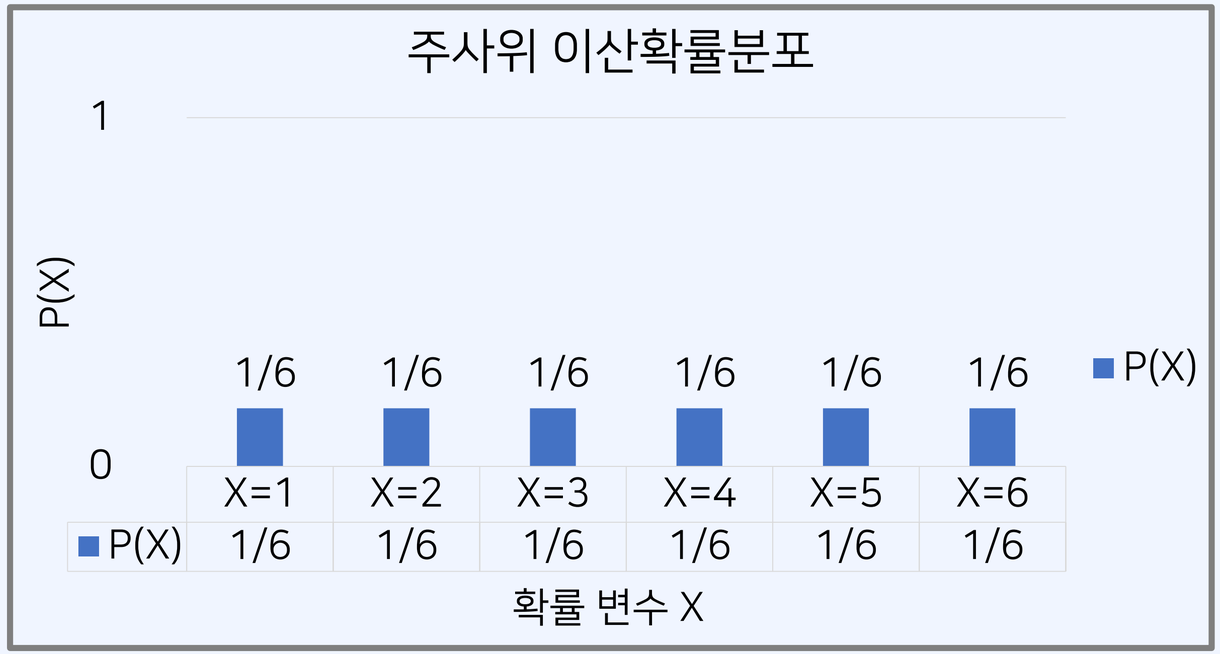
[문제] 확률 변수 X가 구간 $[10, 20]$에서 균등한 분포를 가질 때, $P(15 \leq X \leq 17)$은?

[풀이] X의 확률 밀도 함수가 $f(x) = \frac{1}{20-10} = \frac{1}{10}$이므로, $P(15 \leq X \leq 17) = \frac{1}{10} \times 2 = \frac{1}{5}$이다.



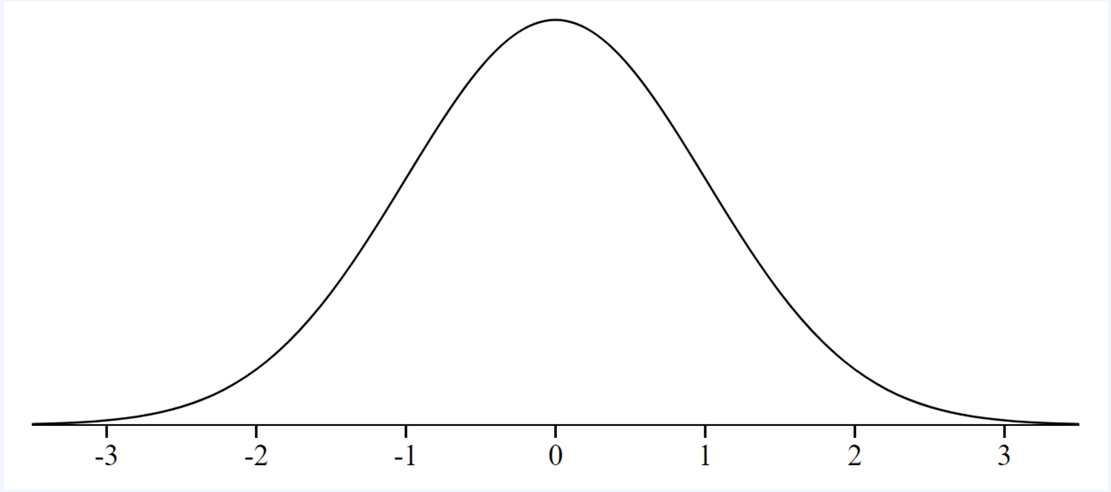
이산 균등 분포(Discrete Uniform Distribution)

- [참고] 균등 분포는 이산확률변수에 대해서도 정의될 수 있다.
- **이산 균등 분포(discrete uniform distribution)**의 예시를 확인해 보자.
- 예를 들어 주사위를 한 번 던지는 시행에서, 주사위의 눈금 값을 확률 변수 X라고 해보자.



정규 분포(Normal Distribution)

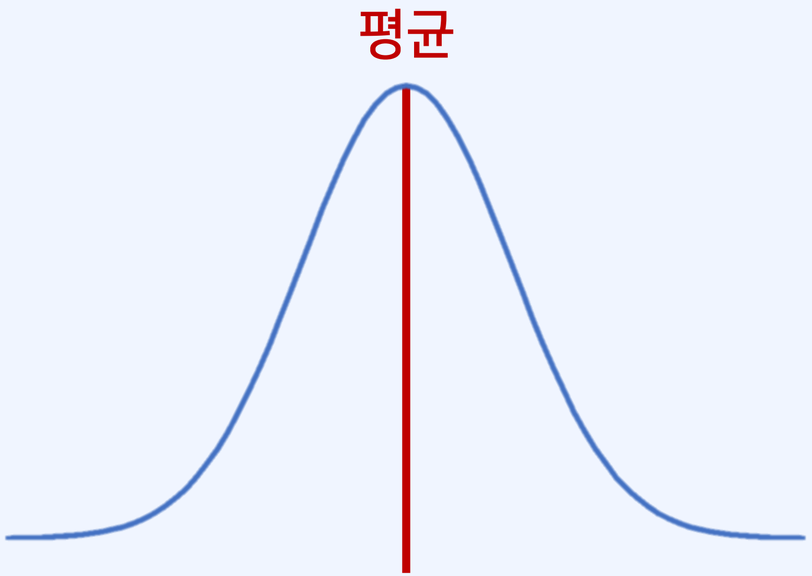
- 정규 분포는 기계학습 분야에서 매우 자주 등장한다.
- 정규 분포는 밥그릇과 같은 모양을 보인다.



* https://www.stapplet.com/normal.html

정규 분포의 특징

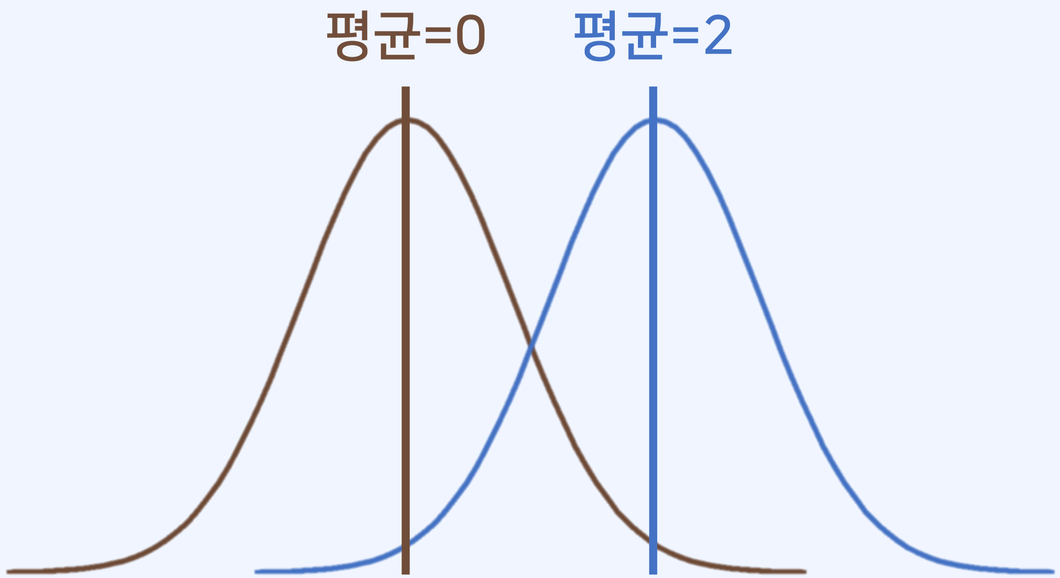
- 정규 분포의 모양은 ① **평균**과 ② **표준편차**로 결정된다.
- 확률밀도함수는 평균을 중심으로 좌우 대칭인 종 모양을 형성한다.
- 관측되는 값의 약 98%가 $\pm 2\sigma$ 범위 안에 속한다.



* https://www.stapplet.com/normal.html

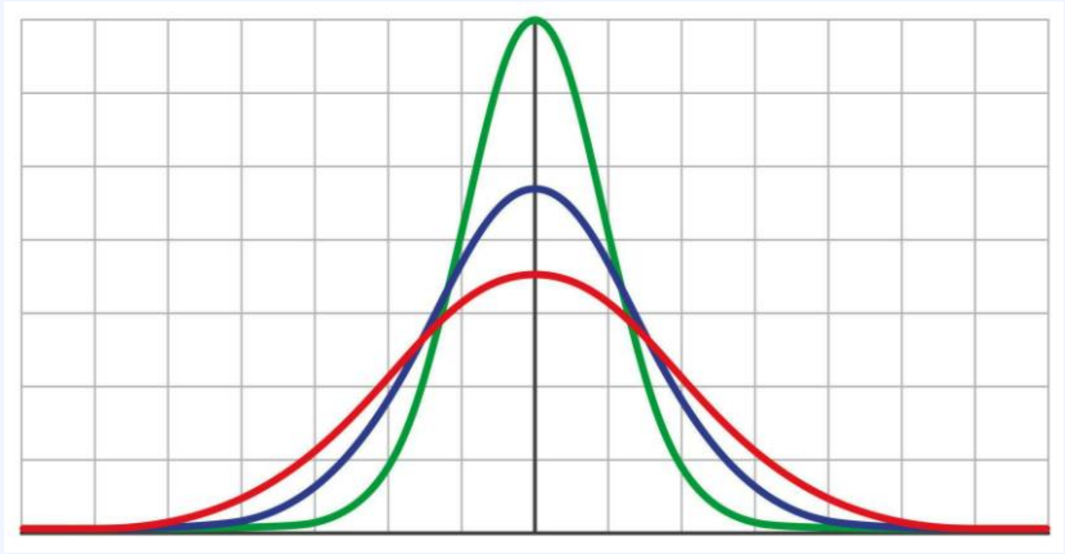
정규 분포의 특징 - 평균

- 평균에 따라서 정규 분포가 좌우로 **평행이동**한다.
- 평균이 0인 정규 분포와, 평균이 2인 정규 분포를 비교할 수 있다.



정규 분포의 특징 - 분산

- **분산이 클수록** 정규 분포가 옆으로 넓게 퍼지게 된다.
- **분산이 작을수록** 정규 분포는 가파른 모양을 가진다.



정규 분포의 확률밀도함수

- 공학 분야에서는 **가우시안(Gaussian) 분포**로 부르기도 한다.
- 확률 변수 X의 확률밀도함수가 다음과 같을 때, X가 정규분포를 따른다고 한다.
- 평균 μ과 분산 σ^2에 의해 분포의 모양이 결정된다. $(-\infty < x < \infty; \sigma > 0)$

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

정규 분포(Normal Distribution)

- 현실 세계의 많은 데이터의 상당수가 **정규 분포로 매우 가깝게 표현**되는 경향이 있다.
- 그래서 현실 세계에서 수집된 데이터의 분포를 근사할 때 자주 사용된다.

표준 정규 분포(Normal Distribution)

- **표준 정규 분포**는 평균이 0, 표준편차가 1인 정규분포를 의미한다.
- 확률을 계산하기 위해 정규 분포 함수를 직접 적분하는 것은 매우 어렵고 번거롭다.
- 실제로는 정규 분포 함수를 **표준 정규 분포로 변환**한 뒤에 **확률을 계산**한다.

지수 분포(Exponential Distribution)

- 특정 시점에서 어떤 사건이 일어날 때까지 걸리는 시간을 측정할 때 사용한다.
- 웹 사이트에 평균적으로 10분에 한 명씩 방문자가 접속한다.

[문제] 한 명의 방문자가 접속한 뒤에, 그 다음 방문자가 올 때까지 걸리는 시간의 확률 구해보자.

- 앞서 포아송 분포는 **발생 횟수**에 대한 확률을 구할 때 사용했다.
- 지수 분포는 **대기 시간**에 대한 확률을 구할 때 사용한다.

	분류	설명
포아송 분포	이산 확률 분포	발생 횟수에 대한 확률
지수 분포	연속 확률 분포	대기 시간에 대한 확률

지수 분포(Exponential Distribution)

- 지수 분포의 확률 밀도 함수는 다음과 같다.
- λ: **단위 시간** 동안 평균 사건 **발생 횟수**

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

지수 분포 예시 – 해킹 시도

[문제]

- 운영 중인 서버에는 하루 평균 4건의 해킹이 시도된다.
- 첫 번째 해킹 시도가 3시간 안에 발생할 확률은?

[해설]

- 24시간에 4건의 해킹이 발생하므로, 단위 시간이 “시간(hour)“일 때 평균 발생 횟수는 4/24다.
- 따라서, $\lambda = 4/24$이며, 확률 밀도 함수는 다음과 같다.

$$f(t) = \frac{4}{24} e^{-\frac{4}{24}t}, t \geq 0$$

지수 분포 예시 – 해킹 시도

[해설]

- 따라서 적분 식을 계산하여 문제를 해결할 수 있다.

$$\int_0^3 \frac{4}{24}\exp\left(-\frac{4}{24}t\right)dt$$

$$= \left[-\exp\left(-\frac{4}{24}t\right)\right]_0^3$$

$$= \exp(0) - \exp\left(-\frac{4}{24}\times 3\right) = 0.3935$$

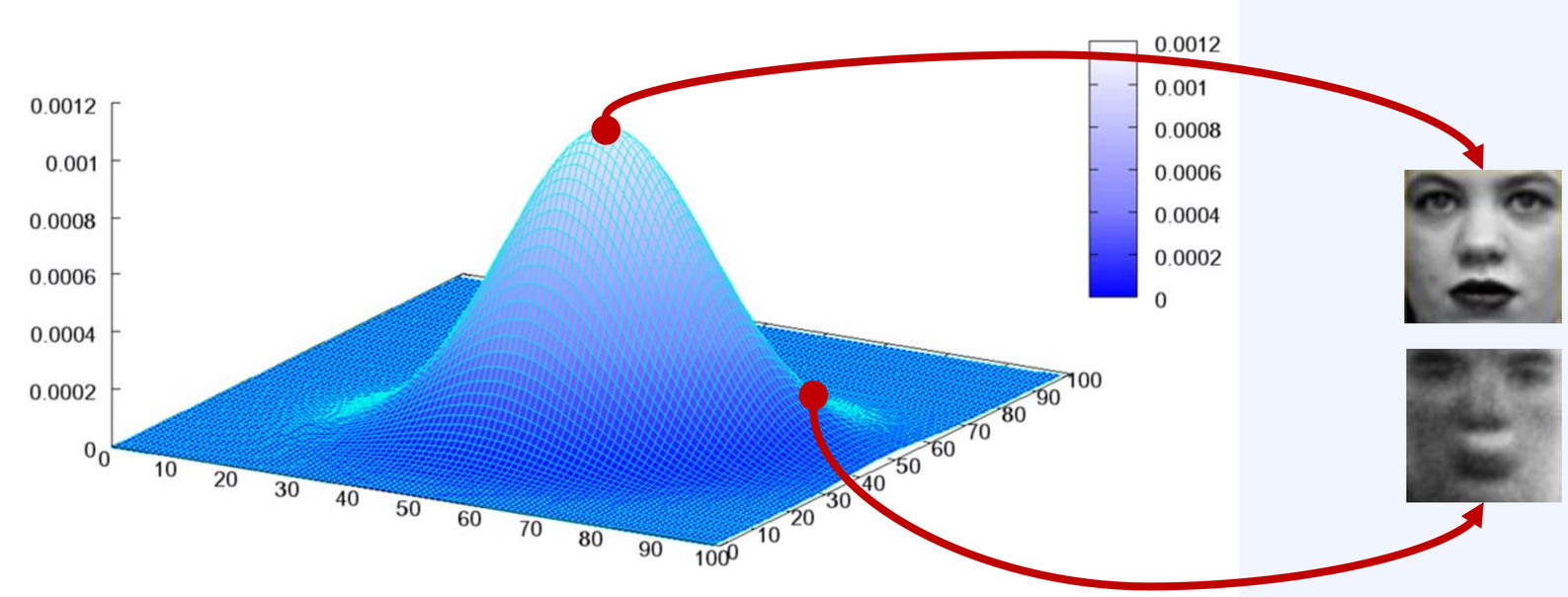
지수 분포의 특성 - 무기억성

- 지수 분포의 특성으로는 **무기억성**이 존재한다.
- 특정 시점에서부터 소요되는 시간은 과거로부터 영향을 받지 않는다.
- 예를 들어, 서버가 해킹당하기까지 걸리는 시간을 지수 분포로 근사한 경우를 고려해 보자.
- 서버를 3년간 운영한 뒤, 해킹당하기까지 걸리는 시간은 처음 서버를 운영한 뒤 해킹당하기까지 걸리는 시간과 같다.

→ **한계점**: 현실 세계에서의 다양한 사례를 모델링하기에는 지나치게 단순한 경향이 있다.

참고: 이미지 데이터에 대한 확률 분포

- **생성 모델**: 이미지의 분포를 근사해, 있을 법한 이미지를 생성할 수 있다.
- 사람의 얼굴에는 통계적인 평균치가 존재할 수 있다.



* https://en.wikipedia.org/wiki/Multivariate_normal_distribution