[CSE4152] 고급소프트웨어 실습 I
Week 5

서강대학교 공과대학 컴퓨터공학과
교수 임 인 성

본 강의에서 제작하여 제공하는 PDF 파일, 동영상, 그리고 예제 코드 등의 강의 자료의 저작권은 특별히 명기되어 있지 않은 한 서강대학교에 있습니다. 본인의 학습 목적 외에 공개된 장소에 올리거나 타인에게 배포하는 등의 행위를 금합니다. 협조 부탁합니다.

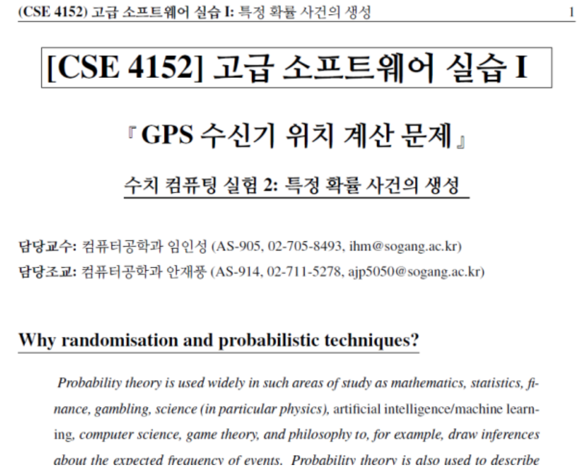
Probability Distribution and Stochastic Event

- **5주차 실습의 목표**
  - **특정 확률 분포(probability distribution)를 따르는 확률 사건(stochastic event)를 수치적으로 충실하게 시뮬레이션**을 해봄.
    - 4주차에 배운 Newton-Raphson 방법과 같은 **비선형 방정식의 수치풀이 방법**
    - Composite trapezoidal rule과 같은 **수치 적분 방법**

- **확률론(probability theory)**
  - 통계학의 수학적 기초로서 확률에 대해 연구하는 수학의 한 분야
    - **확률 변수**(random variable), **확률 과정**(stochastic process), **사건**(event) 등의 내용을 다룸.
  - 비 결정론적(nondeterministic) 현상을 수학적으로 기술함을 목표로 함.

- **응용분야**
  - **무궁무진!**

확률 변수(Random Variable)


(CSE 4152) 고급 소프트웨어 실습 I: 특정 확률 사건의 생성 1

[CSE 4152] 고급 소프트웨어 실습 I

『GPS 수신기 위치 계산 문제』

수치 컴퓨팅 실험 2: 특정 확률 사건의 생성

담당교수: 컴퓨터공학과 임인성 (AS-905, 02-705-8493, ihm@sogang.ac.kr)
담당조교: 컴퓨터공학과 안재풍 (AS-914, 02-711-5278, ajp5050@sogang.ac.kr)

Why randomisation and probabilistic techniques?

Probability theory is used widely in such areas of study as mathematics, statistics, finance, gambling, science (in particular physics), artificial intelligence/machine learning, *computer science, game theory, and philosophy to, for example, draw inferences about the expected frequency of events. Probability theory is also used to describe*

- **확률변수란?**
  - **주어진 범위 내에서 임의의 값을 가질 수 있는 변수**
    - **이산 확률 변수(discrete random variable)**
      - 주사위를 던질 때 나오는 값 X (X = 1, 2, 3, 4, 5, 6)
    - **연속 확률 변수(continuous random variable)**
      - 스마트폰을 새로 사서 고장이 날 때까지의 시간 Y (Y는 양의 실수)

  - 주어진 실험을 통하여 변수 값이 발생함.

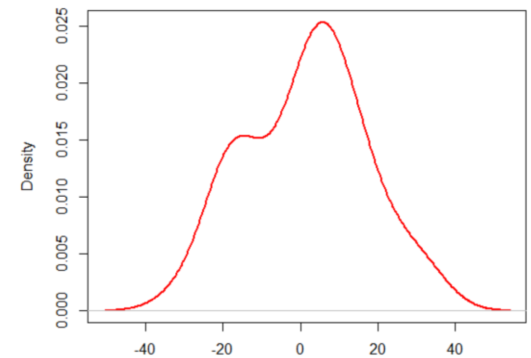
  - **그러한 값이 어떠한 패턴으로 발생하는지를 확률을 통하여 모델링**함.
      - 주사위를 던질 때 5가 나올 가능성은?
      - 스마트폰을 새로 사서 3년 동안 고장 나지 않고 사용할 가능성은?

확률 밀도 함수(Probability Density Function, pdf)

- **정의 (변수 1개)**

  $p_X(x)$를 실수 공간에서 정의된 연속 확률 변수 X의 확률 밀도 함수라 할 때, 이 함수는 다음과 같은 조건을 만족해야 한다.

  1. 모든 실수 값 x에 대해, $p_X(x) \geq 0$
  2. $\int_{-\infty}^{\infty} p_X(x)\, dx \;=\; 1$

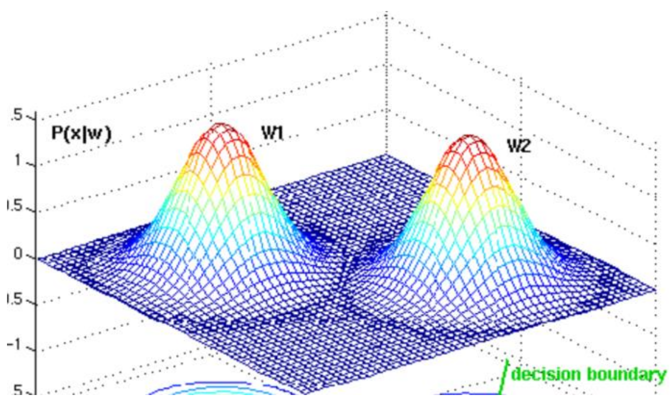


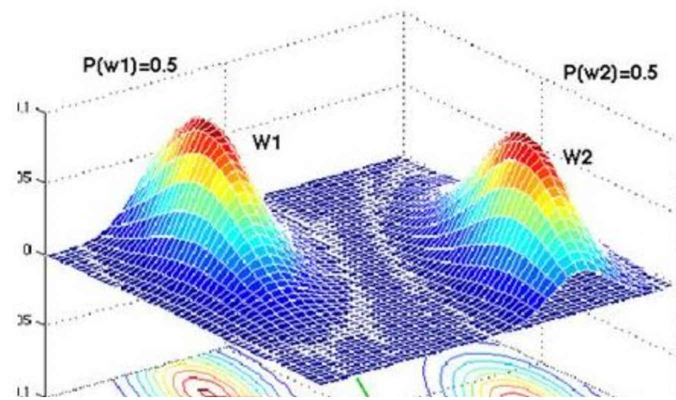
- **확률 밀도 함수와 확률 간의 관계**
  - 확률 밀도 함수는 정확히 말해서 확률값을 나타내는 함수가 아니라 확률 변수 X가 특정 값을 가질 정도를 나타내주는 함수임.
    - 아주 작은 양수 ϵ에 대해, $P(x \leq X \leq x+\epsilon) \;\approx\; p_X(x) \cdot \epsilon$
  - 확률 변수 X가 어떤 범위의 값을 가질 확률은 다음과 같음.
    - 구간 $[a,b]$에 대해 $P(a \leq X \leq b) \;=\; \int_a^b p_X(x)dx$

- **정의 (변수 2개)**

$p_{X,Y}(x)$를 실수 공간에서 정의된 연속 확률 변수 X와 Y의 결합 확률 밀도 함수라 할 때, 이 함수는 다음과 같은 조건을 만족해야 한다.

1. 모든 실수 값 x와 y에 대해, $p_{X,Y}(x,y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dx\,dy \;=\; 1$





누적 분포 함수(Cumulative Distribution Function, cdf)

- **정의**

어떤 확률 변수가 특정 값보다 작거나 같을 확률을 기술해주는 함수. $F_X(x)$를 연속 확률 변수 X의 누적 밀도 함수라 하면,

- $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} p_X(t)\,dt$

- **누적 분포 함수와 확률 밀도 함수와의 관계**

누적 분포 함수의 순간 변화율, 즉 x가 순간적으로 증가 함에 따라 $F_X(x)$, 즉 X가 x보다 같거나 작을 확률이 증가하는 정도가 바로 $p_X(x)$라는 것을 의미함.

- $p_X(x) = \frac{dF_X}{dx}(x)$

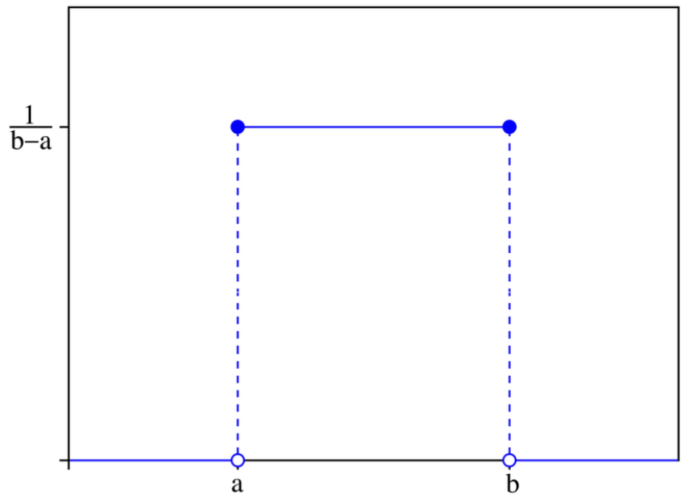
- **누적 분포 함수의 중요한 성질**

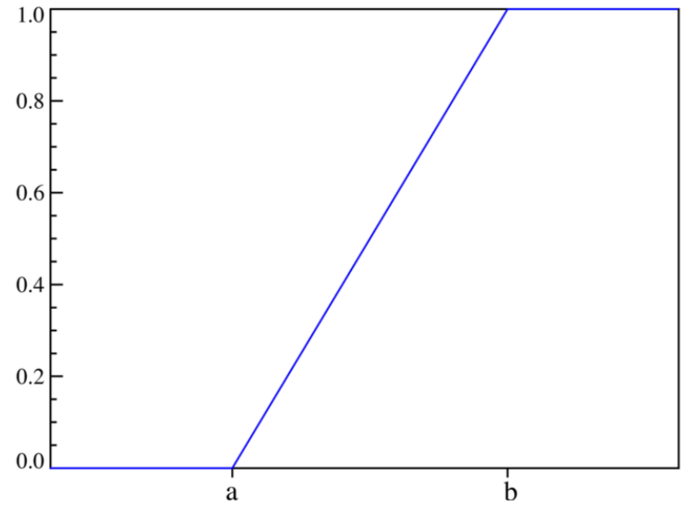
0과 1 사이의 값을 가지는 누적 분포 함수는 단조 비감소 함수 (monotonically nondecreasing function)임.

- 만약 $x_0 \leq x_1$이라면, $F_X(x_0) \leq F_X(x_1)$

- 예: **균등 분포(uniform distribution)**

$$p_X(x) = \begin{cases} 1, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases} \qquad F_X(x) = \int_0^x 1\, dt = x, \quad 0 \le x \le 1$$



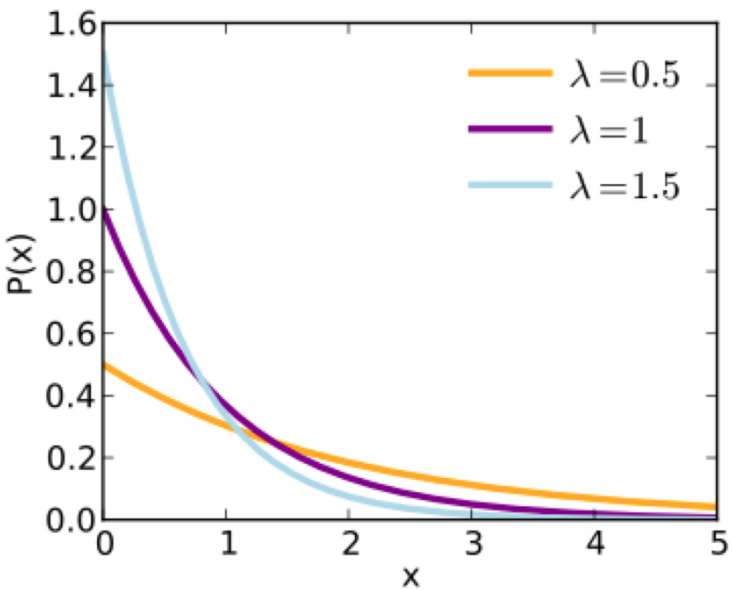


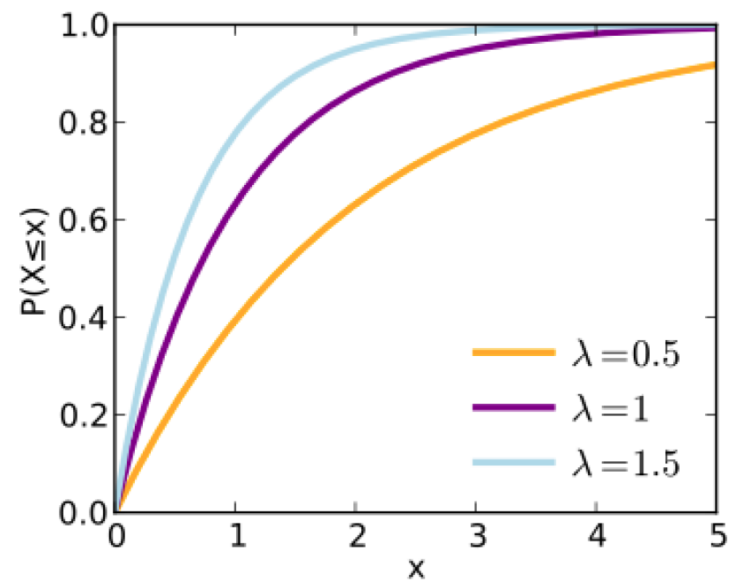
- 예: **지수 분포(exponential distribution)**

$$p_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \qquad F_X(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}, \quad x \geq 0$$

  - To model the time until something happens in the process.
  - Memoryless!

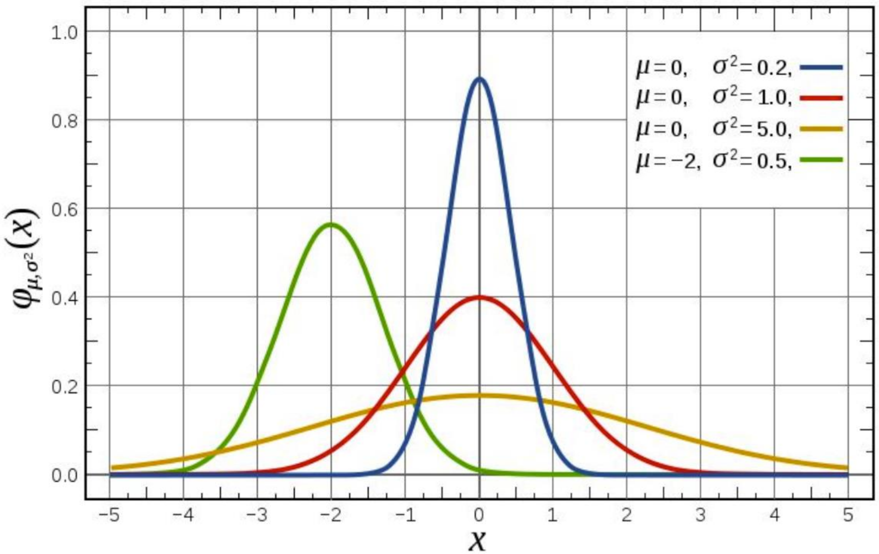
$$P(Y > t \,|\, X > s) = P(X > s + t \,|\, X > s) = \frac{P(X > s + t)}{P(X > s)} = e^{-\lambda t}$$

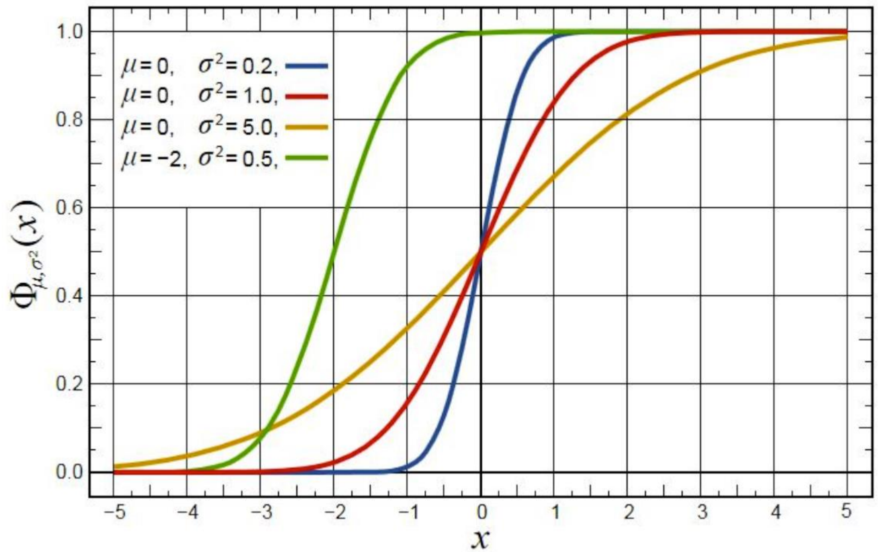




- 예: **정규 분포(normal distribution)**

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad F_X(x) = ???$$



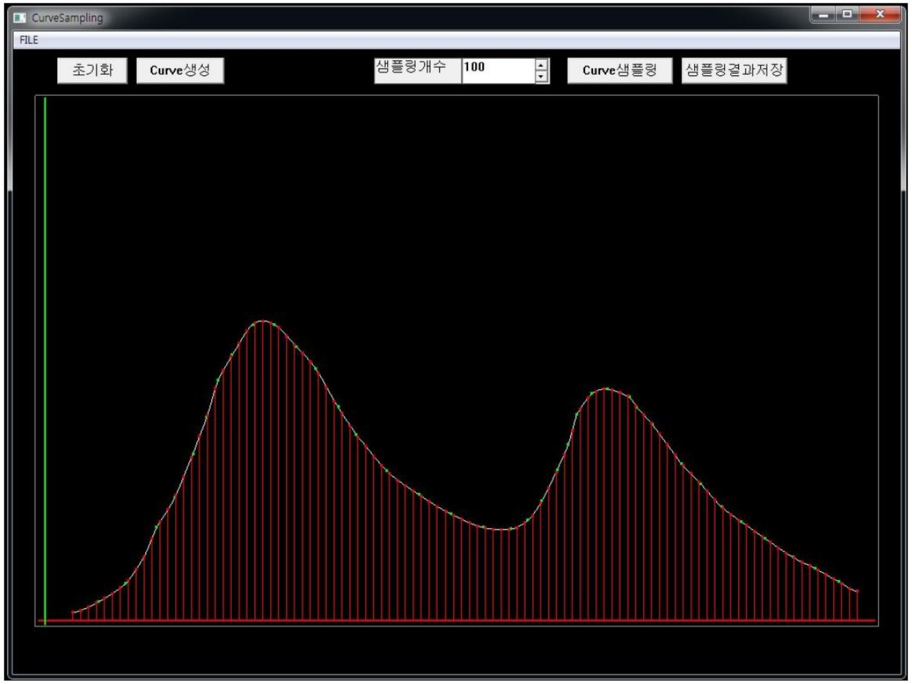


Inversion Method

- **문제**

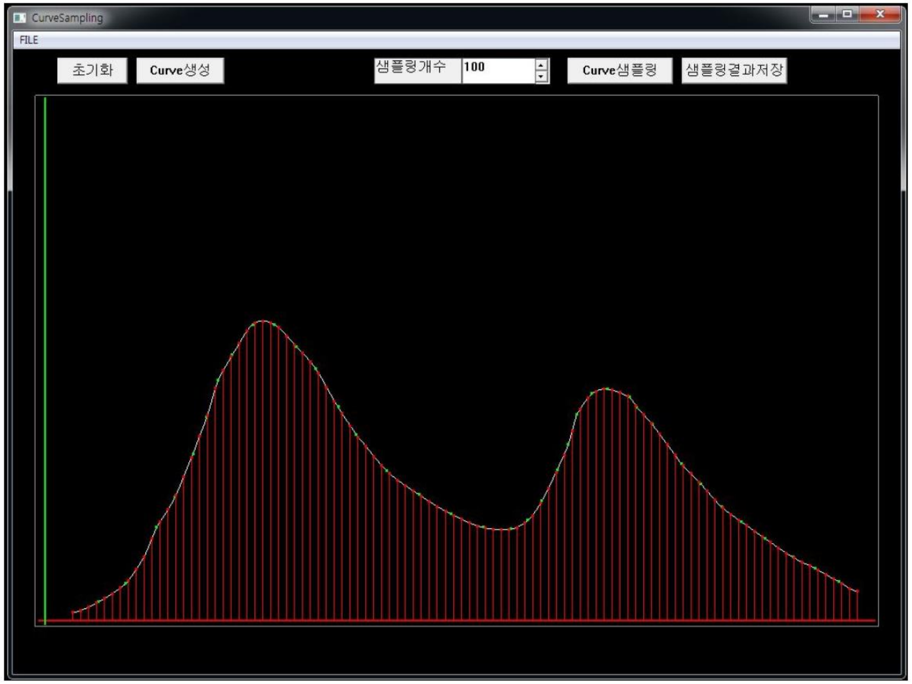
사용자가 임의로 정의한 확률 밀도 함수 $p_X(x)$를 따르는 확률 변수 X를 모사해 주는 난수 수열 $X_0, X_1, X_2, X_3, \cdots$를 발생시켜라.

  - $p_X(x)$를 따르는 난수란 이를 무한히 많이 발생시켜 x값에 대한 분포를 계산할 경우, 그 분포가 $p_X(x)$에 수렴함을 의미함.

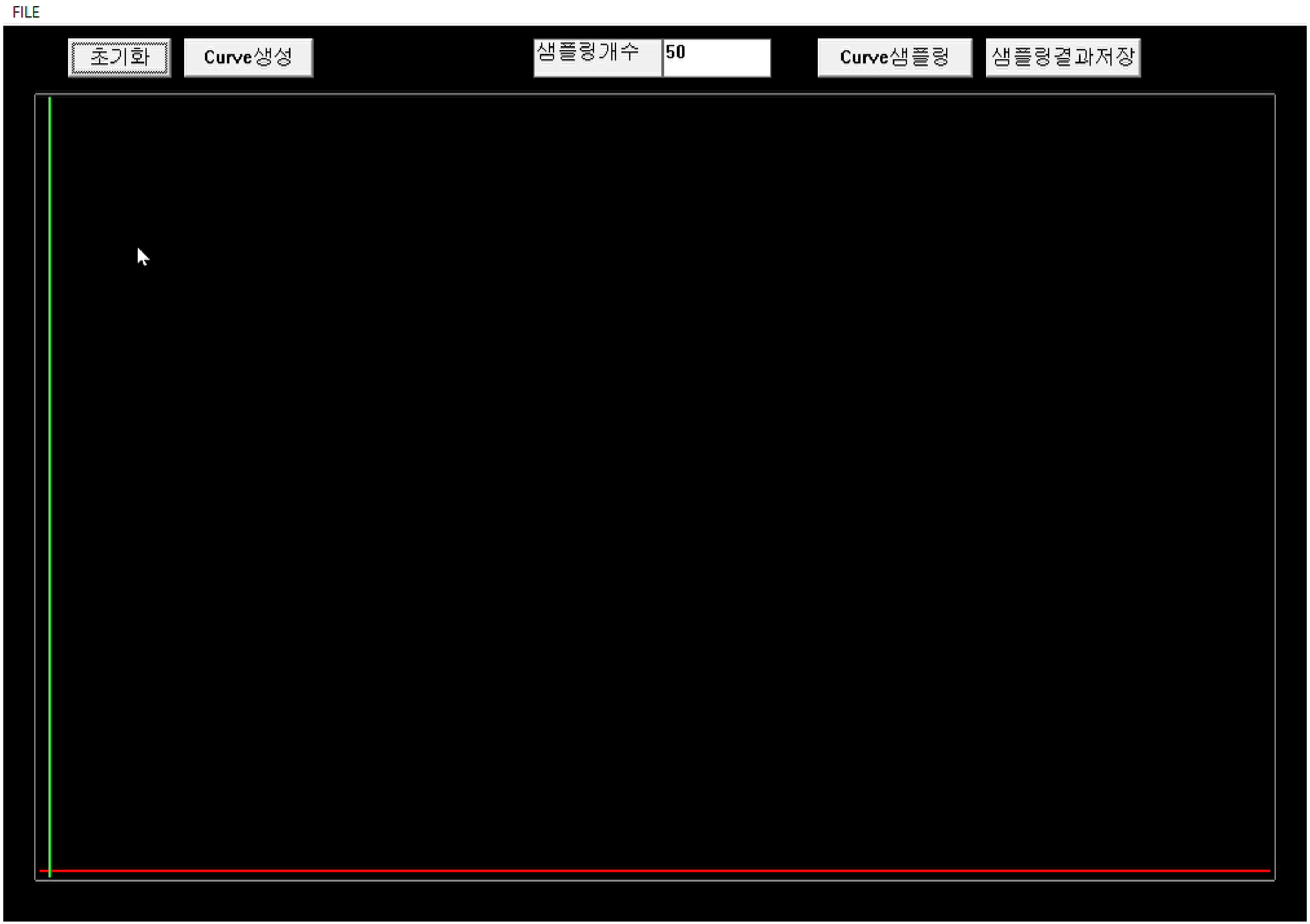



```
100  5.676768
20.000000   9.000000
25.676767   11.591169
31.353535   15.391957
37.030304   20.120052
42.707069   25.132595
      ⋮
564.969727   47.323036
570.646484   42.336571
576.323242   36.826424
582.000000   32.999992
```

- **본 과목 제공 pdf 생성 코드**
  - 아래와 같이 곡선 생성을 통하여 대략적인 그래프를 구한 후, 전체 면적으로 함수 값을 나누어 원하는 pdf를 생성함.

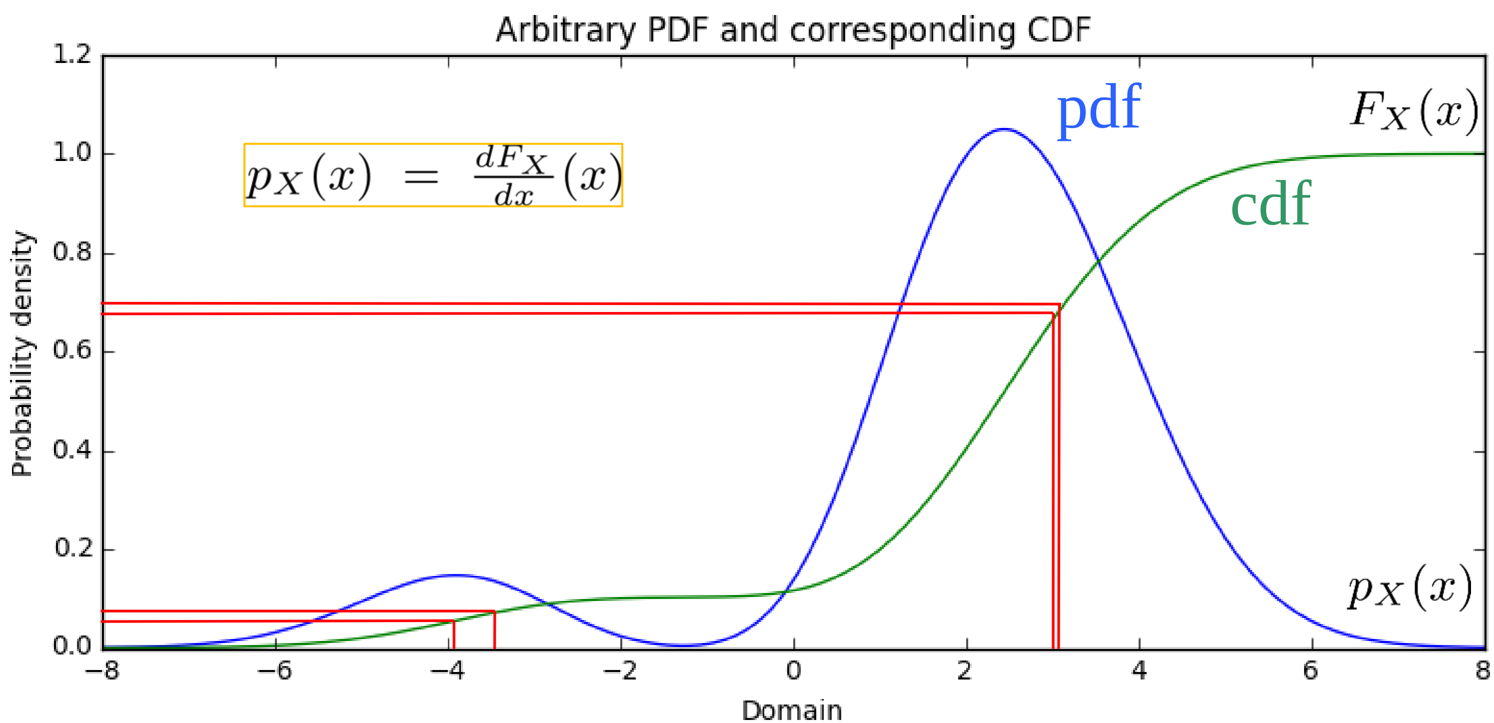



```
100   5.676768
20.000000   9.000000
25.676767   11.591169
31.353535   15.391957
37.030304   20.120052
42.707069   25.132595
        ⋮
564.969727   47.323036
570.646484   42.336571
576.323242   36.826424
582.000000   32.999992
```


FILE
초기화
Curve생성
샘플링개수
50
Curve샘플링
샘플링결과저장

• **PDF와 CDF 간의 관계**

$U_0, U_1, U_2, U_3, \cdots$를 $[0, 1]$ 구간의 값을 가지는 균등 확률 변수에 대하여 생성한 난수 수열이라고 하자. $F_X(x)$를 주어진 확률 밀도 함수 $p_X(x)$에 대한 누적 분포 함수라 할 때, 난수 수열 값 $X_i = F_X^{-1}(U_i)$의 직관적인 의미는?



- **방법**

  $U_0, U_1, U_2, U_3, \cdots$를 $[0, 1]$ 구간의 값을 가지는 균등 확률 변수에 대하여 생성한 난수 수열이라고 하자. $F_X(x)$를 주어진 확률 밀도 함수 $p_X(x)$에 대한 누적 분포 함수라 하면, 이 확률 분포를 따르는 난수 수열 값 X_i $(i = 0, 1, 2, 3, \cdots)$는 $X_i = F_X^{-1}(U_i)$와 같이 구할 수 있다.

$$\boxed{p_X(x) \rightarrow F_X(x) = P(X \leq x) = \int_{-\infty}^{x} p_X(t)\, dt \rightarrow U_i \text{ from } U[0, 1] \rightarrow X_i = F_X^{-1}(U_i)}$$

- Intuitive explanation

$$P(X_i \leq x) = P(F_X^{-1}(U_i) \leq x) = P(U_i \leq F_X(x)) = F_X(x)$$

- 예: 지수 분포

$$U_i = F_X(X_i) = 1 - e^{-\lambda X_i} \longrightarrow X_i = -\frac{\ln\ 1 - U_i}{\lambda}$$

$$\boxed{p_X(x) \rightarrow F_X(x) = P(X \le x) = \int_{-\infty}^{x} p_X(t)\,dt \rightarrow U_i \text{ from } U[0,1] \rightarrow X_i = F_X^{-1}(U_i)}$$

- **몇 가지 문제점**
  - 일반적으로 확률 밀도 함수가 수식으로 표현이 되는 경우도 있지만 이산적인 형태로 근사적으로 주어지기도 함 → 적절한 가공이 필요함.

1. 모든 실수 값 x에 대해, $p_X(x) \ge 0$
2. $\int_{-\infty}^{\infty} p_X(x)\,dx = 1$

$$\longrightarrow \quad p_X(x) = \frac{f(x)}{\int_{x_0}^{x_n} f(x)\,dx}, \quad x \in [x_0, x_n]$$

  - 누적 분포 함수의 역함수를 수식으로 표현하는 것이 불가능하거나 매우 어려움 → **비선형 방정식의 근을 구하는 방식으로 해결 가능**.

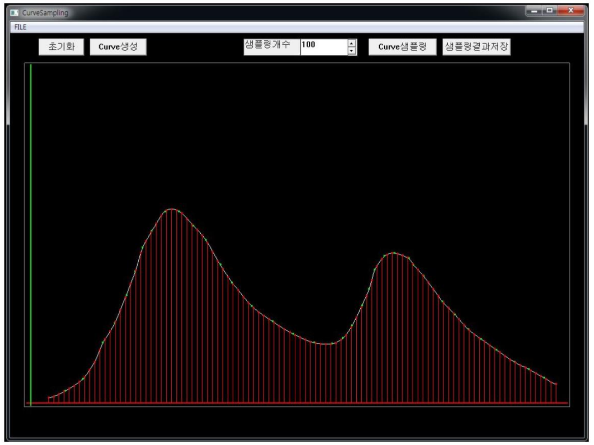
$$\boxed{X_i = F_X^{-1}(U_i) \to F_X(x) = U_i \to f(x) \equiv F_X(x) - U_i = 0, \quad x \in [x_0, x_n] \to X_i}$$

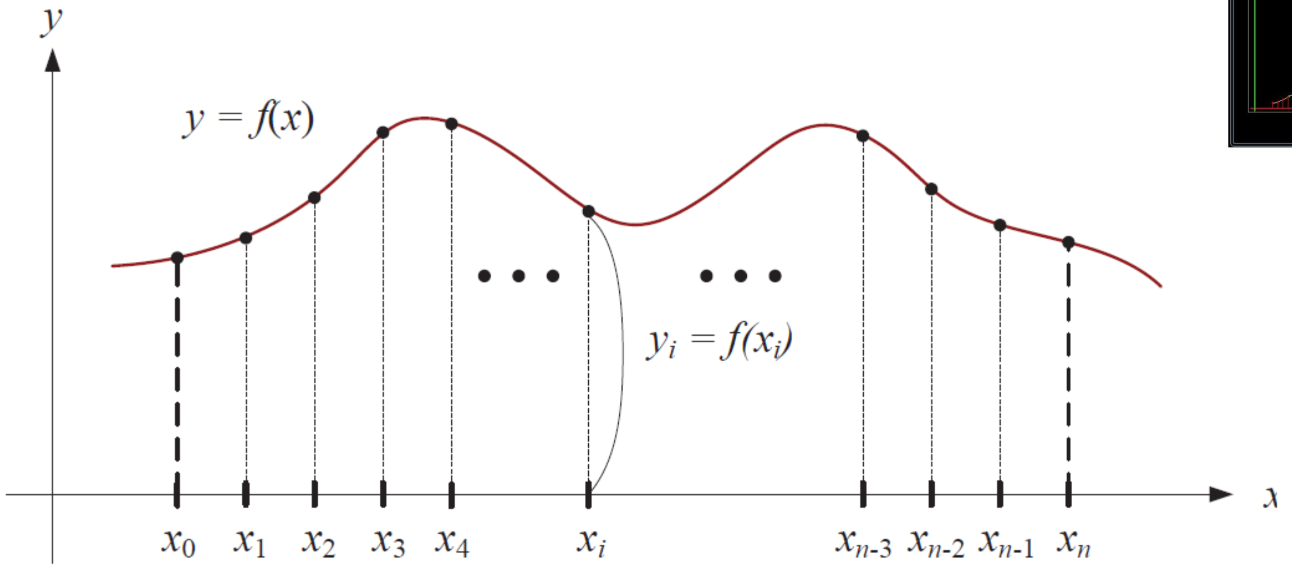
$$\text{Newton iteration: } x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

  - 누적 분포 함수 값을 구하기 위한 적분 계산을 정확한 수식을 사용하는 것이 불가능하거나 매우 어려움 → 수치적인 방법을 사용하여 적분 값을 계산함.

Mathematical Function in Real-life

x	x_0	x_1	x_2	$\cdots$	x_{n-1}	x_n
y	y_0	y_1	y_2	$\cdots$	y_{n-1}	y_n

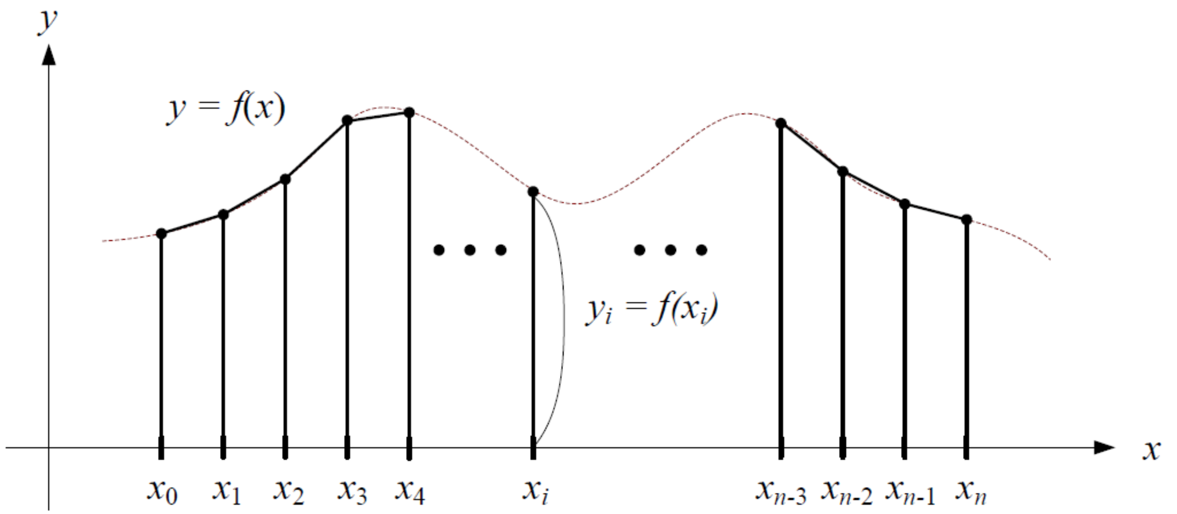




$$\Delta h = \frac{b-a}{n}, \quad x_i = a + i \cdot \Delta h \ (i = 0, 1, 2, \cdots, n), \quad y_i = f(x_i) \ (i = 0, 1, 2, \cdots, n)$$

수치 적분 (Numerical Integration)

- 합성 사다리꼴 공식(composite trapezoidal rule)



$$\int_a^b f(x)\,dx \;\approx\; \frac{\Delta h}{2}\left\{ f(x_0) \;+\; 2\sum_{i=1}^{n-1} f(x_i) \;+\; f(x_n) \right\}$$

$$E = -\frac{f''(\xi)(x_n - x_0)}{12} h^2, \;\; \xi \in [x_0, x_n]$$

$$\Delta h = \frac{b-a}{n}, \quad x_i = a + i \cdot \Delta h \;\; (i = 0, 1, 2, \cdots, n), \quad y_i = f(x_i) \;\; (i = 0, 1, 2, \cdots, n)$$