전산통계

Chapter 2. 확률분포의 생성

분포의 생성

- 분포란 확률변수가 갖는 값들에 확률이 대응되어 있는 것
- 확률분포를 확률분포표로 나타낼 수 있으며, 히스토그램과 같은 그래프로 나타낼 수 있음
- 확률이란 모든 경우를 고려하였을 때 특정한 사건이 발생할 비율
- 확률분포 또한 실험의 모든 결과를 고려할 때 나타나는 것
- 이산형 확률변수, 연속형 확률변수
- 확률분포의 생성은 확률분포를 따르는 확률변수값들을 충분히 생성하였을 때, 의미를 가짐
- 분포가 의미를 가지도록 충분히 많은 확률변수값을 생성하는데 시간이 오래걸리므로 가능한 동일한 효과를 가지면서 시간이 적게 걸리는 생성 알고리즘 개발이 중요

- 원하는 확률 분포를 따르는 확률변수값을 생성하기 위한 단계

  1단계 : 0과 1사이의 균일분포에 따르는 난수들을 생성

  2단계 : 생성된 난수들을 사용하여 원하는 확률변수값 생성 알고리즘에 의해 생성

난수생성자

- 확률변수 X가 균일분포를 따를 때 $X \sim U(0,1)$라고 표현

- $f(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & x \notin [0,1]' \end{cases}$ $F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$
- 1부터 1000까지의 정수값을 갖는 이산형 균일분포를 생성하기 위해서는 서로 구분되지 않는 구슬 1000개을 잘 섞은 후 복원추출하여 난수를 생성

- 순환공식 (recursive formula)
  - 순환공식은 앞의 값의 함수형태로 다음 값을 구하는 결정적 모형식으로 수열을 생성
  - $x_{n+1} = g(x_n), n = 1, 2, \cdots$
  - 난수 (random number)란 균일분포를 따르는 모집단으로부터 추출된 확률표본의 관측값이어야 하므로 순환공식을 이용하는 알고리즘에 의하여 생성된 난수는 유사난수 (pseudo-random number)
  - 생성된 유사난수의 형태와 성질이 실제 난수와 아주 가깝다면 난수의 역할을 한다고 보아도 무방
  - 실제 난수와 차이가 거의 없는 유사난수를 생성하는 전산 알고리즘을 난수생성자 (random number generator)라고 함

- 난수생성자의 조건
  - 생성된 난수들의 분포가 균일분포를 따라야 한다.

    생성된 난수가 초깃값과 생성 개수에 무관하게 균일분포를 따라야 한다.
  - 생성된 자료들 사이에 독립성이 있어야 한다.

    임의로 선택하는 초깃값이 무엇이 되든지 이후 생성되는 난수들 사이에

    자기상관 (autocorrelation)이 없어야 한다.
  - 동일한 난수수열의 재생성이 가능하여야 한다.

    동일한 초깃값을 주면 난수의 수열이 동일하게 재생성되어야 한다.
  - 생성에 효율적이어야 한다.

선형합동법 (linear congruential method)

- $x_{n+1} = a \cdot x_n + c \ (mod\ m),\ n = 0,\ 1,\ 2,\ \cdots$
- $x_0,\ a,\ c,\ m > 0$
- $x_0,\ a,\ c < m$
- 생성 알고리즘

  1단계 : 초기화 단계) $x_0,\ a,\ c,\ m$을 정해준다.

  2단계 : 반복 생성 단계) $x_n = a \cdot x_{n-1} + c \ (mod\ m),\ n = 0,\ 1,\ 2,\ \cdots$

$$u_n = \frac{x_n}{m}$$

  3단계 : 통계적 실험 단계) 생성된 난수열을 통계적 실험에 사용

$$\left\{0,\quad \frac{1}{m},\quad \frac{2}{m},\quad \cdots,\quad \frac{m-1}{m}\right\}$$

- 생선된 난수가 가질 수 있는 값들은 m가지의 값이 나오며 등간격으로 나타남
- 원하는 난수는 연속형 균일분포에 따라야 하므로 가능한 잘게 나누어질수록 좋음 (즉, m이 크면 클수록 좋음)
- 언제나 m가지의 값이 나오는 것은 아니며 $x_0,\ a,\ c$의 값에 따라 나올 수 있는 경우가 달라질 수 있음

- 예) $m = 8$일 때, 다음 두 가지 경우에 대하여 난수의 생성결과 비교

  1) $a = 3,\ c = 1,\ x_0 = 3$

  $$x_n = 3 \cdot x_{n-1} + 1\ (mod\ 8)$$

  3, 2, 7, 6, 3, 2, 7, 6, ….

  주기가 4인 수열

  2) $a = 5,\ c = 3,\ x_0 = 3$

  $$x_n = 5 \cdot x_{n-1} + 3\ (mod\ 8)$$

  3, 2, 5, 4, 7, 6, 1, 0, 3, 2, 5, 4, 7, 6, 1, 0, …

  주기가 8인 수열

- 선형합동법을 이용한 난수생성자의 조건
  - 법수 m의 값이 클수록 좋다.
  - 생성된 난수의 주기는 크면 클수록 좋다.
  - 완전주기

- 완전 주기를 위한 조건
  - c와 m은 공통인수를 가지지 않는 정수여야 한다.
  - a는 m의 모든 인수 p에 대하여 $a = 1\ (mod\ p)$ 이어야 한다.

- a의 선택은 일반적으로 $\sqrt{m}$보다는 크고, $m-\sqrt{m}$보다는 작도록 잡아준다.
- 선형합동법을 가수의 유무에 따라 구분하여

  $c \neq 0$인 경우를 혼합식 합동법

  $c = 0$인 경우를 승산식 합동법
- 승산식 합동법은 $c = 0$인 경우이므로 계산속도는 혼합식 합동법보다 생성시간이 적게 걸림
- 승산식 합동법은 $c = 0$인 경우이므로 c와 m은 공통인수가 없어야 한다는 완전주기의 필요충분조건을 만족하지 못 함.
- 승산식 합동법은 혼합식 합동법보다 많은 연구가 이루어져 왔음.
- 여러 면을 종합하여 볼 때, 승산식 합동법이 혼합식 합동법보다 더 효율적이라고 판단되어 있어 승산식 합동법이 난수생성자의 기본을 이루고 있음.

표준균일분포의 검정

- 카이제곱분포를 이용한 적합도 검정
- 콜모고로프-스미르노프 적합도 검정
- 그래프를 이용한 난수 사이의 독립성 검정
- 런 검정