Lecture 02 - Discrete Distribution (1)

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

국립 서울과학기술대학교 IT정책전문대학원

I. 이산확률변수 - 확률 밀도

분포(Distribution)

- 통계학에서 가장 기본적인 개념
- **항상 동일한 결과가 나오지 않는 현상**을 수치화하려는 노력이다.

이산일양분포 (Discrete Uniform Distribution)

- 동전이나 주사위를 던졌을 때의 현상을 표현하는 분포

  ”정육면체인 주사위를 던지면 1에서 6사이의 숫자 중에서 하나가 나온다. 그리고 여섯 가지 숫자가 나올 확률을 1/6로 모두 동일하다.”

도표로 표현

- 위의 진술을 어떻게 하면 수치화 할 수 있을까? 표로 그린다면? 주사위를 던져서 나올 수 있는 모든 경우(1~6의 숫자)에 대한 각각의 확률을 표현한다면 아래와 같다.

[표 1] 주사위 확률의 도표1

주사위를 던져서 나온 값	1	2	3	4	5	6
위의 값이 나올 확률	1/6	1/6	1/6	1/6	1/6	1/6

[표 1] 주사위 확률의 도표1

주사위를 던져서 나온 값	1	2	3	4	5	6
위의 값이 나올 확률	1/6	1/6	1/6	1/6	1/6	1/6

- [표 1]에 수학기호를 도입해보자.
- X를 **주사위를 던졌을 때의 눈**을 나타내는 **확률변수(random variable)** 이라고 하면, 아래의 [표 2]가 통계학의 일반적인 표현 방식이다.

[표 2] 주사위 확률의 도표2 (도수분포표, probability frequency table)

x	1	2	3	4	5	6
$\mathbb{P}(X=x)$	1/6	1/6	1/6	1/6	1/6	1/6

확률 변수와 [표 2]에 대한 설명

- X는 **주사위를 던졌을 때의 눈**을 표현하는 변수라고 하였다. 이렇게 결과값이 달라질 수 있는 변수를 **확률 변수(Random variable)**라고 한다.
- [표 2]
  - 첫 번째 행의 1, 2, 3, 4, 5, 6은 X가 가질 수 있는 모든 값을 나열한 것이다.
  - 두 번째 행은 각각의 값들에 대한 확률을 표현한 것이다. 즉, $\mathbb{P}(X=1)=1/6, \mathbb{P}(X=2)=1/6, ..., \mathbb{P}(X=6)=1/6$을 표로 표시한 것이다.
  - $\mathbb{P}(X=1)=1/6$는 X가 1이 될 확률, 즉, $X=1$인 사건(event)이 일어날 확률이 1/6과 같다는 뜻이다.
- $\mathbb{P}(\text{some event})$는 "some event"("어떤 사건")가 일어날 확률을 의미한다. 확률이므로 당연히 "some event"가 일어날 확률은 0과 1사이의 값이다.
- [표 2]에서 확률 변수 X는 대문자로, 확률 변수가 가질 수 있는 값 x는 소문자로 표기한 것에 유의하라.
  - 확률 변수는 대문자로 표현한다.
  - 수치는 소문자로 표현한다.
  - 통계학에서는 대소문자를 철저히 구분하며 각각 다른 의미를 가진다.

차트로 표현

- [표 2]를 차트로 표현한다면 어떤 모양을 가질까?
  - 첫 번째 행 x의 수치를 X축으로,
  - 두 번째 행 $\mathbb{P}$의 수치를 Y축으로 표현한다면,
  - 아래와 같은 차트 [그림 1]이 된다.

```
plot(x=1:6, y=rep(1/6, 6), type='p')
```

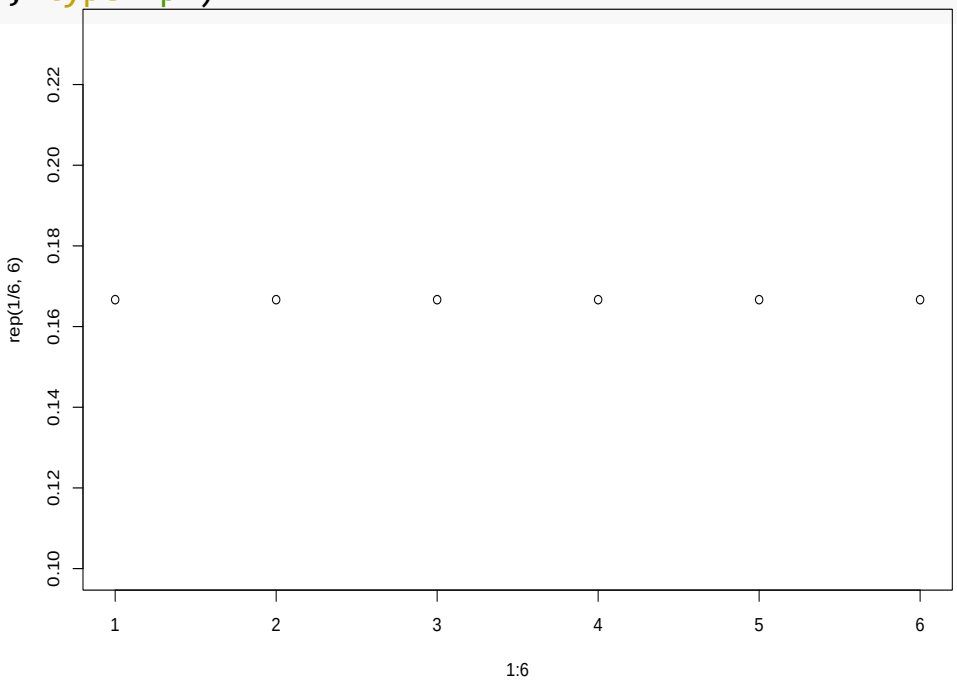

그림 1: 주사위 확률을 표현한 차트1

코드 설명

```
plot(x=1:6, y=rep(1/6, 6), type='p')
```

- `plot()`은 간단한 플랏 기능을 제공하는 R의 함수이다.
- `plot()`의 인수로써
  - `x=1:6`: x값에 1부터 6까지 연속된 정수를 넣는다.
  - `y=rep(1/6,6)`: y값에 1/6을 6번 나열하여 넣는다.
  - `type="p"`는 각각의 `(x,y)`값들을 점(point)로 넣으라는 명령어이다. `"p"`대신에 `"l"`(line), `"b"`(both: point와 line)등의 옵션을 줄 수 있다.
- `plot()`함수에 대해서 더 자세하게 알고 싶으면 `? plot`이나 `help(plot)`을 실행하면 된다.

[그림 1]의 Issues

1. 제목이 없다.
2. 차트의 X 축과 Y 축에 대한 설명이 없다.
3. 각각의 점들이 점이 아닌 원으로 그려져있다.

```
plot(x=1:6, y=rep(1/6,6), type='p',
     main = "Probability for throwing a dice",   # issue 1을 해결
     xlab = "x",                                 # issue 2을 해결
     ylab = "Probability",                       # issue 2을 해결
     pch=16)                                     # issue 3을 해결
```

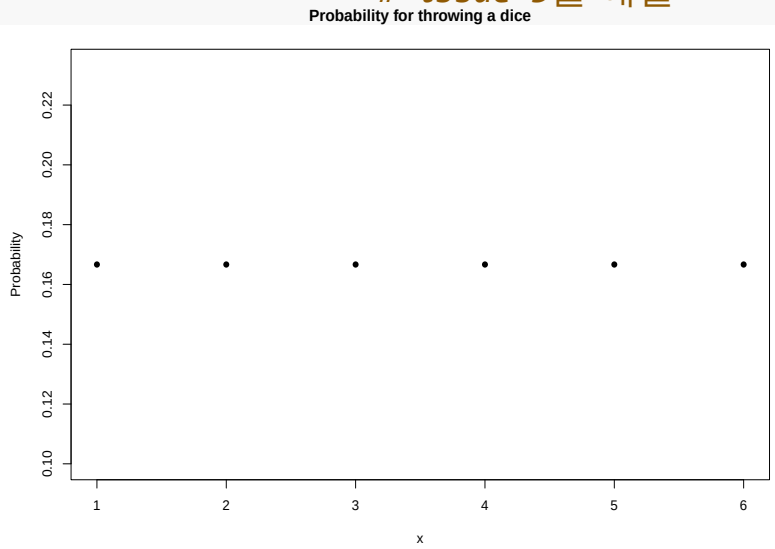

그림 2: 주사위 확률을 표현한 차트2

수식으로 표현

같은 정보를 수식으로 표현한다면?

- 1,2,3,4,5,6을 입력(input)으로 하고,
- 위의 입력에 대해서는 1/6의 출력(output)이 나와야 한다.
- 위의 6개의 입력이 아닌 다른 입력에 대해서는 0이 출력으로 나와야 한다.
- 이를 수식으로 표현하면 아래와 같다.

$$p(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

Some terminologies so far

$$p(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

- 함수 $p(x)$는 $x = 1, 2, 3, 4, 5, 6$과 같은 연속적이지 않은 값들에 대한 확률값을 표현하는 함수이다.
- ($x = 1, 2, 3, 4, 5, 6$는 연속적이지 않은가??)
- Definitions
  - **이산 확률 변수(Discrete random variable)**: 가질 수 있는 값이 연속적이지 않은 확률 변수 (Ex: 현재 예제에서 X)
  - **확률 질량 함수(probability mass function, pmf)**: 이산확률변수가 값 x를 가질 확률을 표현하는 함수 (Ex:현재 예제에서 $p(x)$)

확률 질량 함수(probability mass function, pmf)의 특징

1. 모든 x에 대해서 $0 \leq p(x) \leq 1$이다.
2. 모든 x에 대한 $p(x)$를 더하면 그 값은 1과 같다. 즉, [표 2]에서 두 번째 행에 있는 숫자를 모두 합하면 1이다. 수식으로는 $\sum_{\forall x} p(x) = 1$이다.

Summary so far

- 확률에 관한 정보를 1) 표로 나타내고 2) 차트로 나타내고 3) 수식으로 정리하였다.
- 확률 변수, 이산 확률 변수, 확률 질량 함수에 대해서 정의하였다.
  - 주사위의 눈처럼 확률적으로 달라지는 변수를 확률변수라 하고 X라 적었다.
  - 확률변수 X는 이산적인 값을 취할 수 있기에 이산 확률변수이다.
  - 이산확률변수 X가 가질수 있는 값들에 대한 확률을 표현하는 것이 확률질량함수이다.
  - 이산확률변수의 확률질량함수의 값이 모두 같은 경우에 **이산일양분포(discrete uniform distribution)**를 따른다고 한다.

II. 이산확률변수 - 누적 분포

도표로 표현

누적 분포표 (cumulative probability table)

[표 3] 주사위 확률의 도표3

x	1	2	3	4	5	6
$\mathbb{P}(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
$\mathbb{P}(X \leq x)$	1/6	2/6	3/6	4/6	5/6	6/6

- [표 2]를 확장하여 위와 같이 그리는 것이 가능하다.
- $\mathbb{P}(X \leq x)$는 이산확률변수 X가 x보다 작거나 같은 값을 가질 확률을 표현했다.
- [표 2]와 [표 3]은 같은 정보를 포함하고 있다.

차트로 표현

```
plot(x=1:6, y=1/6*(1:6), type='p',
     main = "Cumulative probability",
     xlab = "x",
     ylab = "Cumulative Probability",
     pch=16)
```

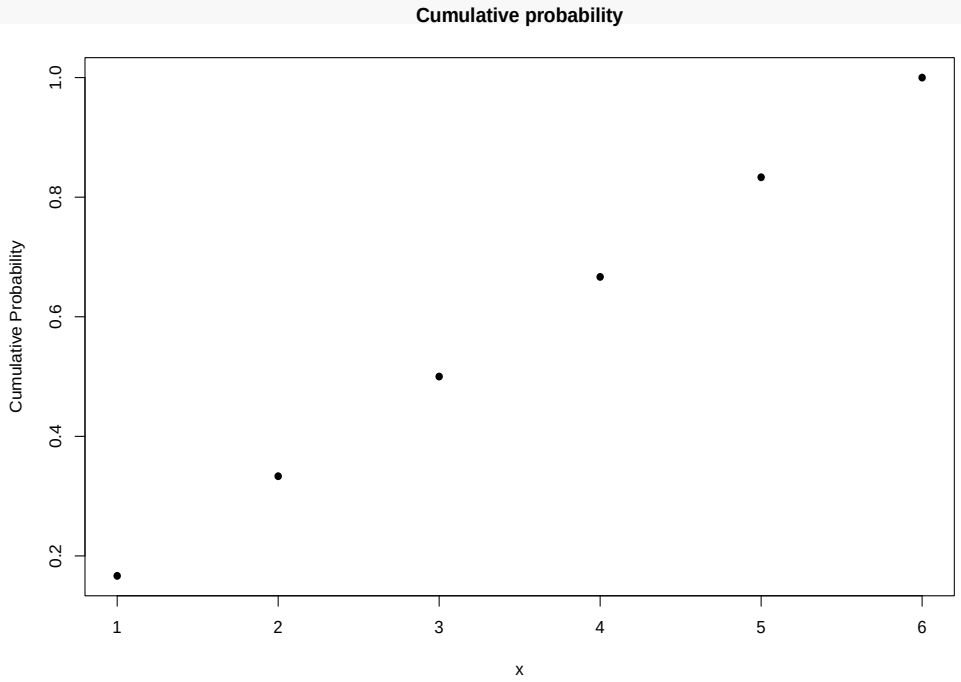

그림 3: 주사위의 누적 확률 차트1

수식으로 표현

- 누적확률에 관한 표현 $\mathbb{P}(X \leq x)$은 x에 관한 함수이며,
- 대문자를 이용해 $F(\cdot)$ 함수로 표기한다.
- 즉, $F(x) = \mathbb{P}(X \leq x)$는 누적확률을 나타내는 함수이고, 이를 **누적분포함수 (cumulative distribution function, cdf)**라고 한다.

More on CDF

- pmf에서는 $x = 1, 2, 3, 4, 5, 6$인 경우에는 $p(x) = 1/6$이고, 다른 x값에 대해서는 모두 $p(x) = 0$이었다.
- cdf에서는 $x = 1, 2, 3, 4, 5, 6$인 경우에는 각각 $F(x) = 1/6, 2/6, 3/6, 4/6, 5/6, 6/6$이 된다.
- 그렇다면 다른 x값에 대해서는 어떻게 될 것인가?
  - $F(1.5)$를 생각해보면, $F(1.5) = \mathbb{P}(X \leq 1.5)$이므로 주사위 눈이 1.5보다 작거나 같은 경우의 확률이다. 주사위 눈이 1.5보다 작은 경우라면 주사위 눈이 1인 경우이다. 즉, $F(1.5) = \mathbb{P}(X \leq 1.5) = \mathbb{P}(X = 1)$이고 1/6과 같다.
  - $F(3.5)$를 생각해보면, $F(3.5) = \mathbb{P}(X \leq 3.5)$이므로 주사위 눈이 3.5보다 작거나 같은 경우의 확률이고, 이는 주사위 눈이 1,2,3중에 하나인 경우이다. 이를 수식으로 표현하면 $F(3.5) = \mathbb{P}(X \leq 3.5) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3)$이고 이는 3/6과 같다.

주사위 눈에 대한 CDF를 아래의 수식으로 정리할 수 있다.

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/6 & \text{if } 1 \leq x < 2 \\ 2/6 & \text{if } 2 \leq x < 3 \\ 3/6 & \text{if } 3 \leq x < 4 \\ 4/6 & \text{if } 4 \leq x < 5 \\ 5/6 & \text{if } 5 \leq x < 6 \\ 6/6 & \text{if } 6 \leq x \end{cases}$$

Issue

- 그렇다면 cdf $F(x)$에 대한 차트 [그림 3]은 제대로 그려진 것인가?
  - [그림 3]은 $x = 1, 2, 3, 4, 5, 6$에 대한 정보만 표시하고 있지만,
  - $x = 1.5$나 $x = 3.5$의 정보는 포함하고 있지 않다.
  - [그림 3]은 수정이 필요하다.

```
plot(x=1:6, y=1/6*(1:6), type='s',
     main = "Cumulative probability",
     xlab = "x",
     ylab = "Cumulative Probability")
points(x=1:6, y=1/6*(1:6), pch=16)
points(x=2:6, y=1/6*(1:6), pch=1)
```

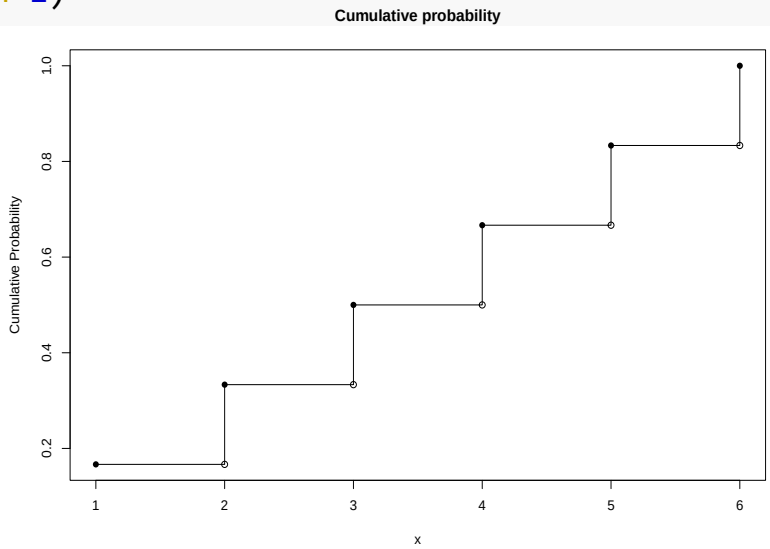

그림 4: 주사위의 누적 확률 차트2

- `plot()`함수의 `type='s'`를 이용하여 계단형(stairstep) 차트를 그린다.
- `points()`함수를 사용하면 앞에서 만들어진 차트위에 점을 표현한다.
- `points()`함수로 색이 칠해진 점과 색이 칠해지지 않은 점을 그려 완성한다.

CDF의 특징

- cdf함수 $F(x) = \mathbb{P}(X \leq x)$는 다음과 같은 특징을 가진다.

1. 확률에 관한 함수이기에 0을 하한, 1을 상한으로 가진다. 즉, $F(x) \geq 0$과 $F(x) \leq 1$이 항상 성립한다.
2. $F(x)$는 감소하지 않는 함수이다. 즉, $x_1 > x_2$인 경우에 항상 $F(x_1) \geq F(x_2)$이다.
3. $F(\infty) = 1$이고 $F(-\infty) = 0$이다.

Summary

- 주사위의 눈을 이용해서 이산일양분포에 대해서 논의하였다.
- 표, 수식, 차트로 표현하면서 확률 변수의 분포에 대한 정보를 정리하는 것은 확률 변수를 이해하는 데 있어서 기본적인 과정이다.

III. Exercises

Problem 1

- 주사위에 대한 논의를 앞면과 뒷면이 나올 확률이 같은 동전(fair coin)에 대해서 반복하라.
- 던진 동전의 면을 X라는 확률 변수로 표현을 하자. 동전을 던져서 앞면이 나온 경우를 1이라고 하고, 뒷면이 나온 경우를 2라고 하자.
  1. [표 2]와 [그림 2]를 그려라.
  2. pmf $p(x)$를 수식으로 표현하라.
  3. [표 3]과 [그림 4]를 그려라.
  4. cdf $F(x)$를 수식으로 표현하라.