6.4 확률분포함수

확률이 어디에 어느정도 분포되어 있는가를 수학적으로 명시하고, 명확하게 전달하기 위한 도구가 바로 확률분포함수다. 이 절에서는 확률분포함수를 정의하는 방법과 확률질량함수, 누적분포함수, 확률밀도함수의 개념을 공부한다.

① 확률분포

$\boxed{\text{확률}}$ = 사건(event)이라는 표본의 집합에 대해 숫자를 할당하는 함수

$\boxed{\text{확률분포}}$ = 어떤 사건에 어느 정도의 확률이 할당되었는지 묘사한 정보를 확률분포(probability distribution)라고 한다.

* 사건을 하나하나 제시하고 할당된 숫자 보여주어야 함.
* 표본 개수가 무한하다면, 모든 사건을 하나하나 기술하는 것은 불가능.
* → 확률분포함수 (probability distribution function)
  - 확률질량 함수 — finite domain
  - 누적분포 함수 } infinite domain
  - 확률밀도 함수 }

〈단순사건과 확률질량함수〉

콜모고로프의 정리를 사용하면 어떤 사건의 확률값을 이용하여 다른 사건의 확률값을 계산할 수 있다. 예를 들어 표본이 하나인 단순사건 (elementary event, atomic event) 이라고 한다.

· 단순사건끼리는 서로 교집합을 갖지 않으므로 유한개의 사건만 있는 경우, 모든 단순사건의 확률값을 알면 콜모고로프의 세 번째 공리에 의해 다른 모든 사건의 확률값을 계산할 수 있다.

*단, 모든 단순사건의 확률 합은 1이어야 한다.

$P(A \cup B) = P(A) + P(B)$
when $P(A \cap B) = \varnothing$

예를 들어, 플레잉카드 무늬 문제의 단순사건과 그 확률이 다음과 같이 정의되어 있다고 하자.

$$P(\{\spadesuit\}) = 0.1, \quad P(\{\heartsuit\}) = 0.2, \quad P(\{\diamondsuit\}) = 0.3,$$
$$P(\{\clubsuit\}) = 0.4$$

다음처럼 모든 사건에 대한 확률을 계산할 수 있다.

$$P(\{\heartsuit, \diamondsuit\}) = 0.2 + 0.3 = 0.5$$

이렇게 유한 개의 사건이 존재하는 경우 각 단순사건에 대한 확률만 정의하는 함수를 확률질량함수(probability mass function)라고 한다.
확률질량함수는 소문자 p로 표시한다. 확률과 확률질량함수는 다른 개념이라는 점을 주의한다.

$$p(a) = P(\{a\})$$

예를 들어 원소가 하나뿐인 사건 {1}에 대한 확률은 확률함수로 정의할 수 있다.

$$p(\{1\}) = 0.2$$

같은 내용을 확률질량함수로 나타내면 다음과 같다.

$$p(1) = 0.2$$

하지만 확률함수가 원소 2개 이상인 사건에 대해서도 확률을 정의할 수 있는데 반해,

$$P(\{1, 2\}) = 0.3$$

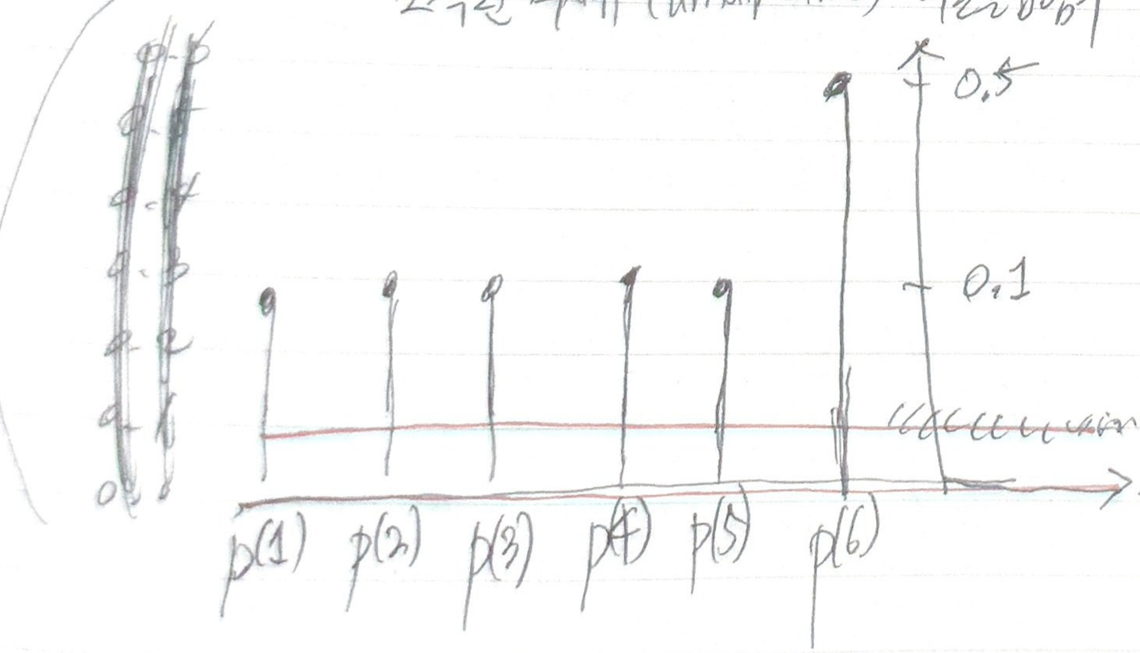
확률질량함수는 사건이 아닌 원소 (정확히 말하면 그 원소만을 가진 단순사건)에 대해서만 정의되므로, 다음과 같은 식은 틀린 식이다.

$$p(1, 2)$$

예제

다음 확률질량함수는 주사위 눈금 6이 다른 숫자보다 비정상적으로 많이 나오게 조작된 주사위 (unfair dice)를 묘사한다.

조작된 주사위 (unfair dice) 확률질량함수



NO.

DATE.

연습문제 6.4.1

확률질량함수가 위와 같은 주사위에서 다음 사건에 대한 확률을 구하라.

(1) $\{1, 2\}$

$$P(\{1,2\}) = P(\{1\}) + P(\{2\})$$
$$= 0.1 + 0.1 = 0.2$$

(2) $\{4, 5, 6\}$

$$P(\{4,5,6\}) = P(\{4\}) + P(\{5\}) + P(\{6\})$$
$$= 0.7$$

$\boxed{\text{표본 수가 무한한 경우}}$

확률질량함수에서 표본공간에 있는 표본수가 유한할 때 하나하나의 표본에 대해서만 확률을 정의하면 어떠한 사건에 대해서도 확률을 정의할 수 있다는 것을 알았다.

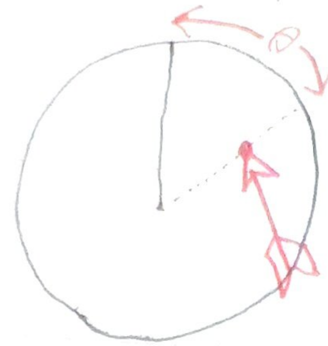
→ 그런데 왜 굳이 확률을 정의할 때 입력을 표본이 아닌 사건으로 정의했을까?

⇒ 표본공간에 있는 표본수가 무한한 경우를 다루기 위해서.

(표본 수가 무한하면 확률질량 함수를 사용하여 확률을 정의할 수 없다)

〈예제〉

다음 그림과 같이 회전하는 원반에 화살을 쏘고 화살이 박힌 위치의 각도를 결정하는 문제를 생각해보자. 각도가 정확하게 0도가 될 확률은 얼마일까?



- 모든 각도에 대해 가능성이 똑같다면, 각도가 정확하게 θ이 될 확률은 0 이다.

$$P(\{\theta = 0^\circ\}) = 0$$

$$P(\{\theta = 30^\circ\}) = 0$$

∴ 표본이 무한하기 때문에.

그렇다면, $P(\{0^\circ \le \theta \le 30^\circ\}) = ?$

모든 가능성이 동일하다면, $\frac{1}{12}$ (같은 조각원반이 12개)

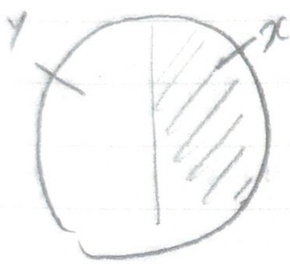
$$P(\{0^\circ \le \theta < 30^\circ\}) = \frac{1}{12}$$

$$P(\{30^\circ \le \theta < 60^\circ\}) = \frac{1}{12}$$

$$P(\{0^\circ \le \theta < 60^\circ \text{ or } 90^\circ \le \theta < 150^\circ\}) = \frac{1}{3}$$

연습문제 6.4.2

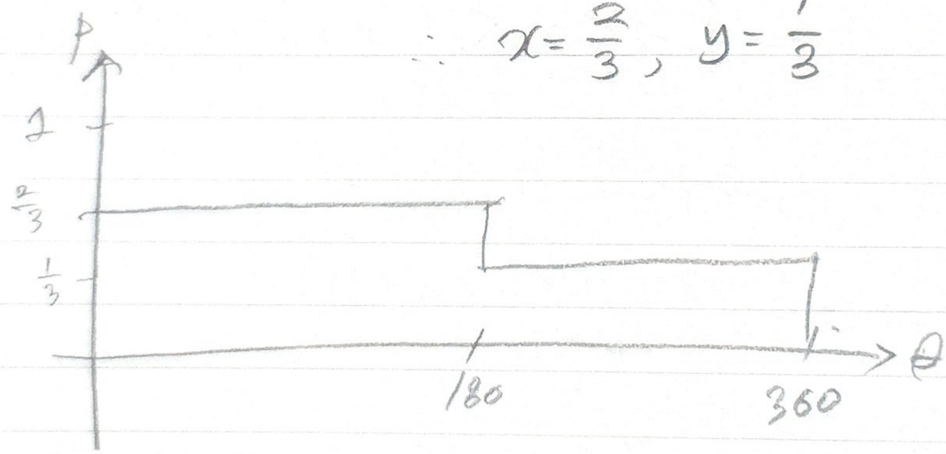
위 예제의 원반을 이용하여 복권 번호를 결정하는 경우를 생각하자.
결과를 조작하려고 0도에서 180도 사이에 화살이 2배 더 잘 박히도록
원반을 조작했다. 이 결과를 확률을 사용하여 공범에게 알려줘야 한다.
가능한 모든 사건에 대해 확률을 알려주는 확률함수를 기술하는 방법은 무엇인가?



$$P(\{0 \le \theta < 180\}) = x$$
$$P(\{180 \le \theta < 360\}) = y$$

① $x+y=1$
② $x=2y$

$\therefore x=\frac{2}{3},\ y=\frac{1}{3}$



구간.

표본공간이 실수의 집합이라면 대부분의 사건 (부분집합)은 시작점과 끝점이라는 두 숫자로 이루어진 구간 (interval)으로 표현된다.

$$A = \{a \leq x \leq b\}$$

$$P(A) = P(\{a < x \leq b\}) = P(a, b)$$

* 구간의 확률만 표현할 수 있다면, 여러 구간으로 이루어진 복잡한 사건은 콜모고로프의 공리에 따라 각 구간의 확률값에 더하거나 빼기로 표현할 수 있다.

예를 들어 다음과 같은 사건

$$B = \{\underline{-2 < x \leq 1} \text{ or } \underline{2 < x \leq 3}\}$$

의 확률 $P(B)$는 다음 두 구간의 확률의 합이다.

$$P(B) = P(\{-2 < x \leq 1\}) + P(\{2 < x \leq 3\})$$

$$= P(-2, 1) + P(2, 3)$$

누적분포함수. CDF

사건 (event) 즉, 구간 (interval) 하나를 정의하기 위해 숫자가 하나가 아닌 두 개가 필요하다는 점은 아무래도 불편하다. 숫자 하나만으로 사건, 즉 구간을 정의할 수 있는 방법은 없을까? 이를 해결하기 위한 아이디어는, 시작점을 모두 똑같이 음의 무한대 ($-\infty$)로 통일한 특수한 구간 S_x을 사용하는 것이다.

$$S_{-1} = \{-\infty < X \le -1\}$$

$$S_0 = \{-\infty < X \le 0\}$$

$$S_1 = \{-\infty < X \le 1\}$$

$$\vdots$$

$$S_x = \{-\infty < X \le x\}.$$

이러한 사건의 확률분포를 묘사하는 함수를 누적분포함수 (cumulative distribution function) 라고 하고, 약자로 cdf 라고 쓴다. 함수기호로는 $F(x)$ 등 대문자 기호로 표시한다. 독립변수 x는 구간의 끝점을 뜻한다.

$$F(x) = P(S_x) = P(\{X < x\})$$

모든 실수는 당연히 $-\infty$보다 크기 때문에, $-\infty$ 부분은 생략한다.

누적분포 함수와 콜모고로프의 공리 3

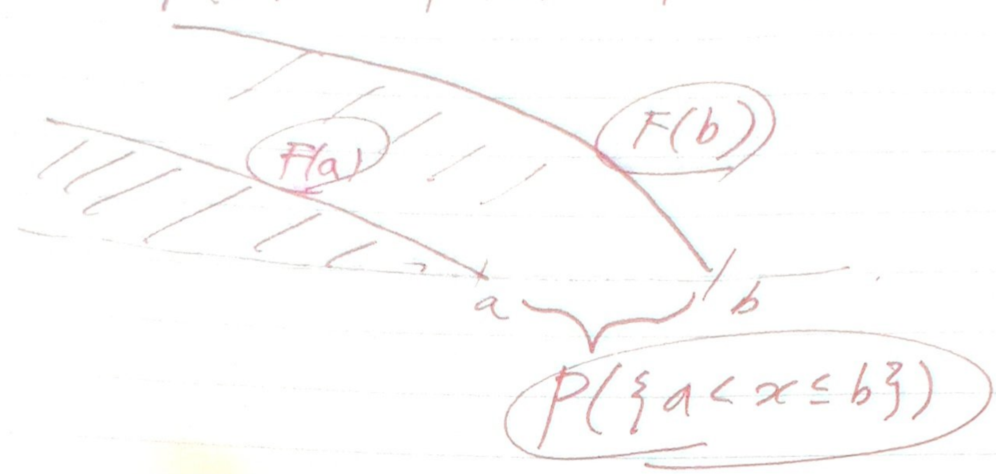
$$A \cap B = \emptyset \longrightarrow P(A \cup B) = P(A) + P(B)$$

를 이용하면, 이러한 사건 S_x의 확률값으로부터 대부분의 복잡한 구간 혹은 사건에 대한 확률값을 계산할 수 있다.

$$P(-\infty, b) = P(-\infty, a) + P(a, b)$$

이를 누적분포함수로 표현하면,

$$F(b) = F(a) + P(a,b)$$



$$F(b) - F(a) = P(\{a < x \le b\})$$
$$= P(a,b)$$

누적분포함수 (cdf)의 특징 :

① 음의 무한대에 대한 누적분포함수값은 0이다.

$F(-\infty) = 0$ $\quad P(\{-\infty < x \le -\infty\})$

② 양의 무한대에 대한 누적분포함수값은 1이다.

$F(+\infty) = 1$ $\quad P(\{-\infty < x \le +\infty\})$

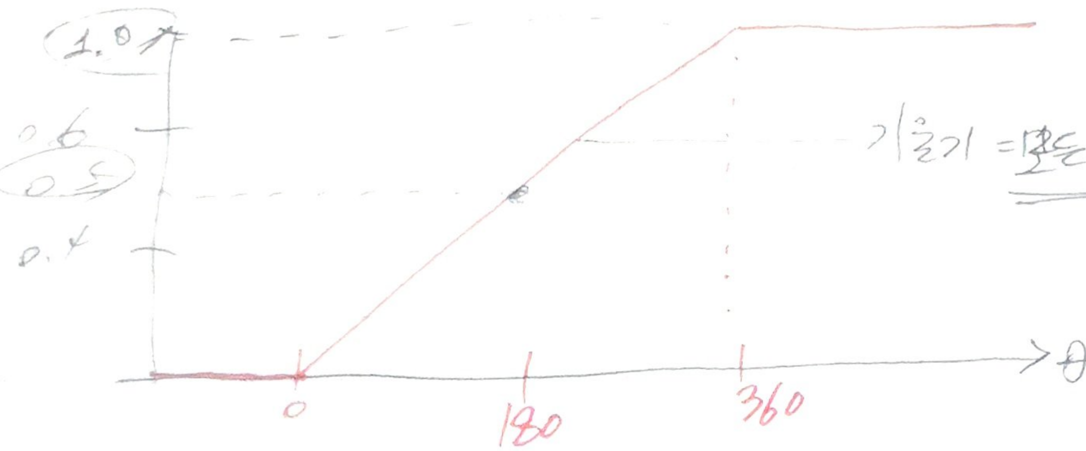
③ 입력이 크면 누적분포 함수값은 커진다 (단조증가)

$$x > y \longrightarrow F(x) \ge F(y)$$

입력값이 커지면 출력값도 항상 증가한다 = 단조증가.

세 가지 특성에 따라, 누적분포함수는 0에서 시작하여 천천히 증가하면서 1로 다가가는 형태를 가진다. 단조증가 성질에 의해 절대 내려가지는 않는다.

ex 원반던지기 누적분포함수 (CDF)



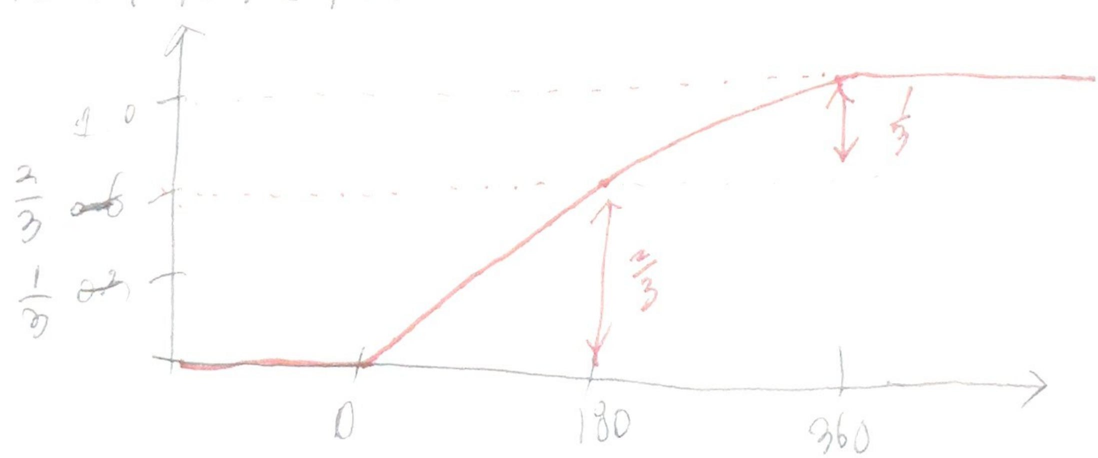
기울기 = 모든 구간에서 동일

✓ 단조증가!
✓ 0부터 1까지 상승!

$$F(180) - F(90) = \frac{1}{2} - \frac{1}{4}$$
$$= P(\{90 < x \le 180\})$$
$$= \frac{1}{4}.$$

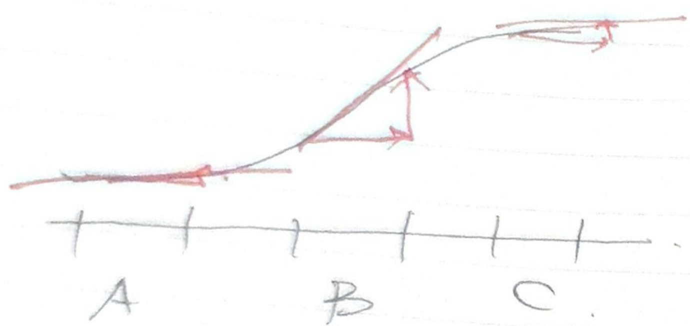
연습문제 6.4.4

0°–180° 사이에 잘 박히도록 조작된 원반을 이용하여 복권번호를 결정하는 문제에서, 누적분포함수 $F(x)$를 구하라.



*CDF의 단점

— 확률이 어느 구간에 많이 할당되었는지 알기 불편함!



기울기!

= 확률이 할당된 정도.

= derivative
(도함수)

⑪ 확률밀도함수 (probability density function)

기울기를 구하는 수학적 연산이 미분 (differentiation)이므로,
누적분포함수를 미분하여 누적분포함수의 기울기를 출력하는 함수를 만들면
어떤 x_1값 근처의 확률이 다른 x_2값 근처보다 더 확률이 높은지, 또는 낮은지
쉽게 파악할 수 있다. 누적분포함수를 미분하여 구한 도함수를 확률밀도함수 (PDF)
라고 한다. 확률질량함수와 마찬가지로 $p(x)$로 표기한다.
(PMF)

$$p(x) = \frac{dF(x)}{dx}$$

derivative이기 때문!

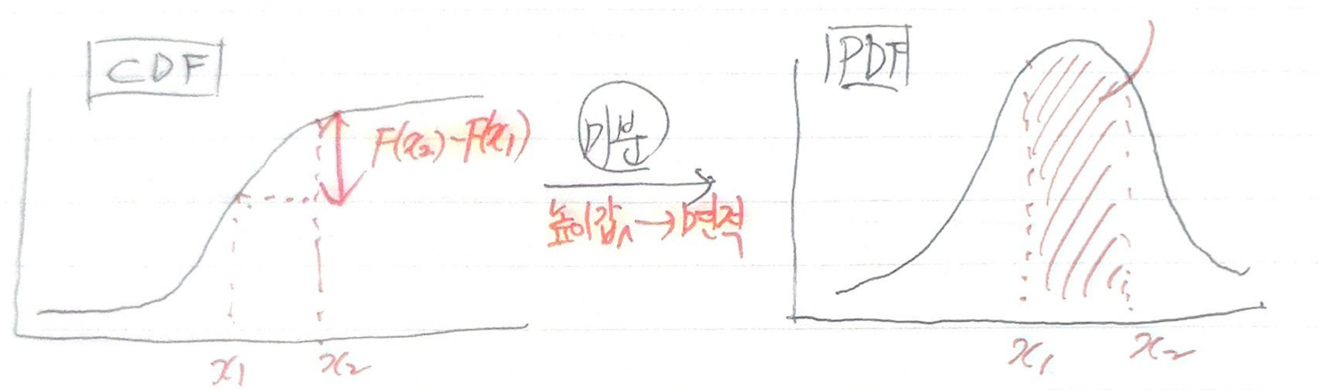
확률밀도함수는 특정한 구간의 확률이 다른 구간에 비해 상대적으로 얼마나 높은가를
나타내는 것이며, 그 값 자체가 확률은 아니라는 점을 명심해야 한다.
함수의 변수가 x가 아니라 u가 된 이유는 x가 적분의 상한인수 (upper bound
argument)로 사용되고 있기 때문이다.

NO.

DATE.

미적분학의 기본 원리에 의하면 $x = x_1$부터 $x = x_2$ 사이에서 도함수인 확률밀도함수의 면적 (정적분)은 적분함수인 누적분포함수의 값을 이용하여 구할 수 있다.

$$F(x_2) - F(x_1) = \int_{x_1}^{x_2} p(u)\,du$$



$$F(x) = \int_{-\infty}^{x} p(u)\,du$$

• 확률밀도함수는 다음과 같은 특징을 갖는다.

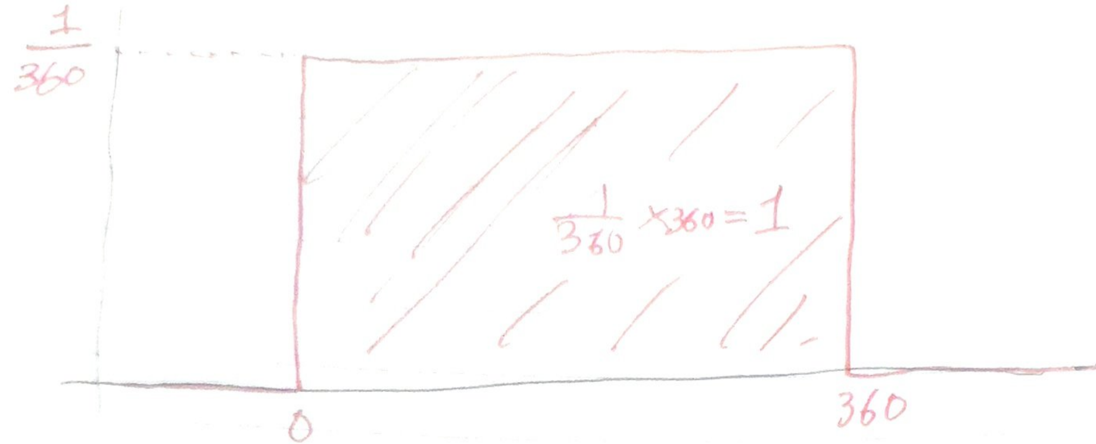
* 적분함수인 누적분포함수의 기울기가 음수가 될 수 없기 때문에, 확률밀도함수는 0보다 같거나 크다.

$$p(x) \geq 0$$

* $-\infty$부터 ∞까지 적분하면, 표본공간 $(-\infty, \infty)$의 확률이 되므로 값은 1이다.

$$\int_{-\infty}^{\infty} p(u)\,du = 1$$

(예제) 공평한 원반에 화살 맞추기 PDF



연습문제 6.4.5

0도에서 180도 사이에 화살이 2배 더 잘 박히도록 조작된 원반을 이용하여 복권 번호를 결정하는 문제에서 확률분포함수 $p(x)$를 구하라.