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.

확률분포

확률은 사건(event)이라는 표본의 집합에 대해 숫자를 할당하는 함수다. 어떤 사건에 어느 정도의 확률이 할당되었는지 묘사한 정보를 **확률분포(probability distribution)**라고 한다. 확률분포를 묘사하려면 모든 사건들을 일일히 제시하고 거기에 할당된 숫자를 보여주어야 한다. 표본의 개수가 유한하다면 가능할 수 있지만 만약 표본의 개수가 무한하다면 현실적으로 모든 사건을 하나 하나 기술하는 것은 불가능하다. 이 절에서는 확률분포함수(probability distribution function)라는 것을 이용하여 이 문제를 해결하는 방법을 설명한다. 확률분포함수로는 다음과 같은 세 종류가 있다.

- 확률질량함수
- 누적분포함수
- 확률밀도함수

단순사건과 확률질량함수

콜로고로프의 정리를 사용하면 어떤 사건의 확률값을 이용하여 다른 사건의 확률값을 계산할 수 있다. 예를 들어 표본이 하나인 사건을 **단순사건(elementary event, atomic event)**이라고 한다. 단순사건끼리는 서로 교집합을 가지지 않으므로 유한 개의 사건만 있는 경우, 모든 단순사건의 확률값을 알면 콜모고로프의 세 번째 공리에 의해 다른 모든 사건의 확률값을 계산할 수 있다. 단 모든 단순사건의 확률의 합은 1이어야 한다.

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

$$P(\{\spadesuit\}) = 0.1, \;\; P(\{\heartsuit\}) = 0.2, \;\; P(\{\diamondsuit\}) = 0.3, \;\; P(\{\clubsuit\}) = 0.4 \tag{6.4.1}$$

다음처럼 모든 사건에 대한 확률을 계산할 수 있다.

$$P(\{\heartsuit, \diamondsuit\}) = 0.2 + 0.3 = 0.5 \tag{6.4.2}$$

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.

$$p(a) = P(\{a\}) \tag{6.4.3}$$

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

$$P(\{1\}) = 0.2 \tag{6.4.4}$$

같은 내용을 확률질량함수로 나타내면 다음과 같다.

$$p(1) = 0.2 \tag{6.4.5}$$

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

$$P(\{1, 2\}) = 0.3 \tag{6.4.6}$$

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

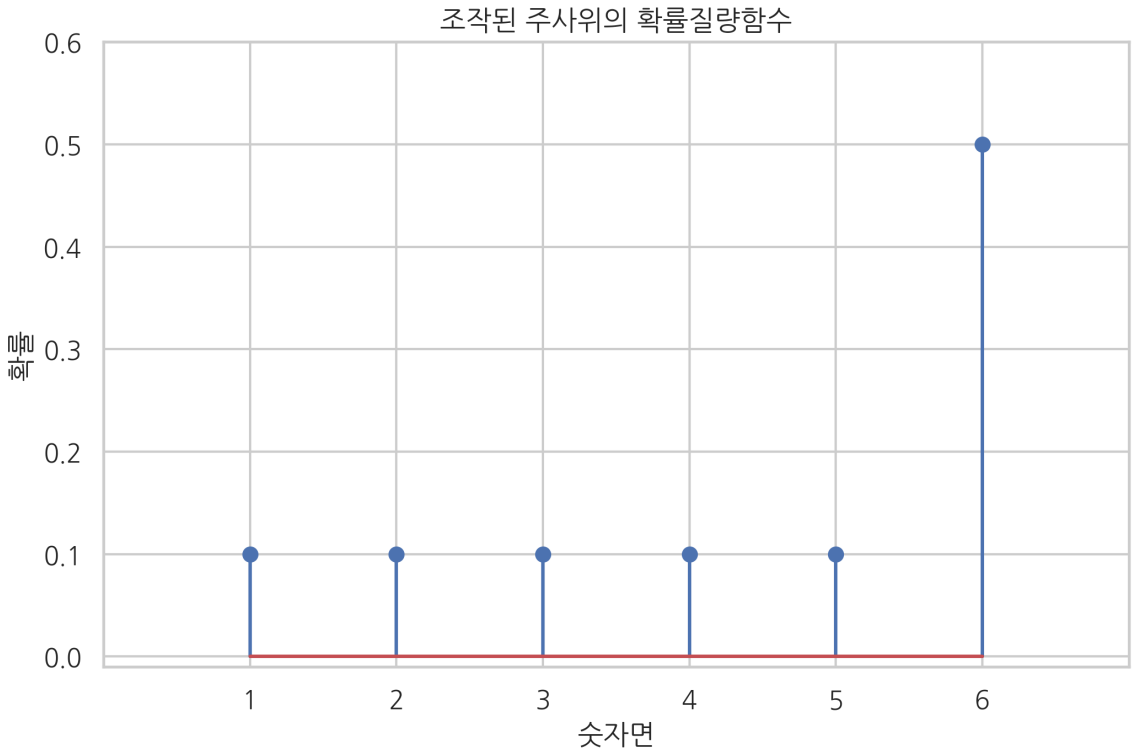
$$p(1, 2) \tag{6.4.7}$$

예제

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

In [1]:

```
x = np.arange(1, 7)
y = np.array([0.1, 0.1, 0.1, 0.1, 0.1, 0.5])
plt.stem(x, y)
plt.title("조작된 주사위의 확률질량함수")
plt.xlabel("숫자면")
plt.ylabel("확률")
plt.xlim(0, 7)
plt.ylim(-0.01, 0.6)
plt.xticks(np.arange(6) + 1)
plt.show()
```



연습 문제 6.4.1

확률질량함수가 위와 같은 주사위에서 다음 사건에 대한 확률을 구하라.

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

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

표본 수가 무한한 경우

확률질량함수에서 표본공간에 있는 표본 수가 유한할 때 하나하나의 표본에 대해서만 확률을 정의하면 어떠한 사건에 대해서도 확률을 정의할 수 있다는 것을 알았다. 그렇다면 왜 굳이 확률을 정의할 때 입력을 표본이 아닌 사건으로 정의했을까? 그 이유는 표본공간에 있는 표본 수가 무한한 경우를 다루기 위해서다. 표본 수가 무한하면 확률질량함수를 사용하여 확률을 정의할 수 없다. 다음 예제를 통해 그 이유를 알아보자.

예제

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

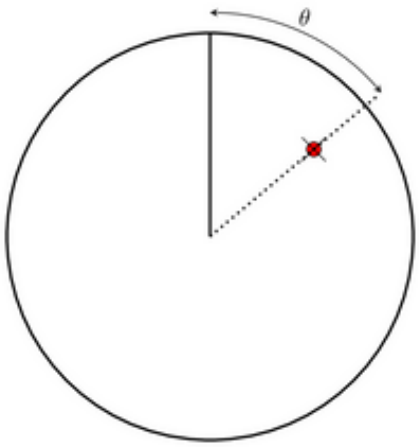


그림 6.4.1 : 회전하는 원반의 각도 문제

만약 모든 각도에 대해 가능성이 똑같다면 **각도가 정확하게 0이 될 확률은 0이다.**

$$P(\{\theta = 0^\circ\}) = 0 \tag{6.4.8}$$

각도가 0이 아닌 어떤 경우도 마찬가지로 확률이 0이다. 예를 들어 각도가 30도가 되는 경우도 확률은 0이다.

$$P(\{\theta = 30^\circ\}) = 0 \tag{6.4.9}$$

왜 그럴까? 모든 각도에 대해 가능성이 똑같으므로 그 확률을 x라는 값이라고 하자. 그런데 각도가 나올 수 있는 경우는 무한대의 경우가 있으므로 만약 x가 0이 아니라면 $x \times \infty = \infty$로 전체 표본 집합의 확률이 무한대가 된다. 즉, 1이 아니다. 따라서 **표본 수가 무한하고 모든 표본에 대해 표본 하나만을 가진 사건의 확률이 동일하다면, 표본 하나에 대한 사건의 확률은 언제나 0이다.**

이번에는 같은 원반에 대해 다음 사건의 확률은 얼마일까?

- 각도가 0도보다 같거나 크고 30도보다 작은 경우

$$P(\{0^\circ \leq \theta < 30^\circ\}) = ? \tag{6.4.10}$$

이 경우에는 동일한 가능성을 지닌 사건이 12개 있으므로 전체집합의 확률 1을 12로 나누면 주어진 사건에 대한 확률은 $1/12$가 된다.

$$P(\{0^\circ \leq \theta < 30^\circ\}) = 1 \div 12 = \frac{1}{12} \tag{6.4.11}$$

다음은 일부 사건에 대해 확률을 할당한 예다.

$$P(\{0^\circ \leq \theta < 30^\circ\}) = \frac{1}{12} \tag{6.4.12}$$

$$P(\{30^\circ \leq \theta < 60^\circ\}) = \frac{1}{12} \tag{6.4.13}$$

$$P(\{0^\circ \leq \theta < 60^\circ \text{ or } 90^\circ \leq \theta < 150^\circ\}) = \frac{1}{3} \tag{6.4.14}$$

이 예제로부터 표본공간의 표본 수가 무한하면 사건에 대해 직접 확률을 할당하는 수밖에 없다는 것을 알 수 있다.

연습 문제 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 전달해야 한다. 가능한 모든 사건에 대해 확률을 알려주는 확률함수를 기술하는 방법은 무엇인가?

구간

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

$$A = \{a < x \leq b\} \tag{6.4.15}$$

a는 구간의 시작점이고 b는 구간의 끝점이다.

구간을 입력받아 확률값을 출력하는 함수는 다음처럼 이차원 함수 $P(a, b)$로 표현할 수 있다.

$$P(A) = P(\{a < x \leq b\}) = P(a, b) \tag{6.4.16}$$

구간의 확률만 표현할 수 있다면 여러 구간으로 이루어진 복잡한 사건은 콜모고로프의 공리에 따라 각 구간의 확률값의 더하기나 빼기로 표현할 수 있다.

예를 들어 다음과 같은 사건

$$B = \{-2 < x \leq 1 \text{ or } 2 < x \leq 3\} \tag{6.4.17}$$

의 확률 $P(B)$는 다음 두 구간의 확률의 합이다.

$$P(B) = P(\{-2 < x \leq 1\}) + P(\{2 < x \leq 3\}) = P(-2, 1) + P(2, 3) \tag{6.4.18}$$

연습 문제 6.4.3

0도에서 180도 사이에 화살이 2배 더 잘 박히도록 조작된 원반을 이용하여 복권 번호를 결정하는 문제에서 구간의 시작점과 끝점을 입력받아서 확률을 출력하는 함수 $P(a,b)$를 구하고 이를 파이썬으로 구현하라.

누적분포함수

그러나 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.

$$\begin{aligned} S_{-1} &= \{-\infty < X \leq -1\} \\ S_0 &= \{-\infty < X \leq 0\} \\ S_1 &= \{-\infty < X \leq 1\} \\ &\vdots \\ S_x &= \{-\infty < X \leq x\} \end{aligned} \tag{6.4.19}$$

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

$$F(x) = P(S_x) = P(\{X < x\}) \tag{6.4.20}$$

모든 실수는 당연히 $-\infty$보다 크기 때문에 $-\infty <$ 부분은 생략했다.

누적분포함수와 콜모고로프의 세 번째 공리

$$A \cap B = \emptyset \quad \rightarrow \quad P(A \cup B) = P(A) + P(B) \tag{6.4.21}$$

를 이용하면 이러한 사건 S_x의 확률값으로부터 대부분의 복잡한 구간사건에 대한 확률값을 계산할 수 있다. 예를 들어 $\{a < x \leq b\}$라는 구간사건의 확률은 콜모고로프의 공리에서

$$P(-\infty, b) = P(-\infty, a) + P(a, b) \tag{6.4.22}$$

로 나타난다. 이를 누적분포함수로 표현하면

$$F(b) = F(a) + P(a, b) \tag{6.4.23}$$

정리하면 다음과 같다.

$$P(a, b) = F(b) - F(a) \tag{6.4.24}$$

누적분포함수 cdf는 다음과 같은 특징이 있다.

- 음의 무한대에 대한 누적분포함수값은 0이다.

$$F(-\infty) = 0 \tag{6.4.25}$$

- 양의 무한대에 대한 누적분포함수값은 1이다.

$$F(+\infty) = 1 \tag{6.4.26}$$

- 입력이 크면 누적분포함수값은 같거나 커진다.

$$x > y \quad \rightarrow \quad F(x) \geq F(y) \tag{6.4.27}$$

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

예제

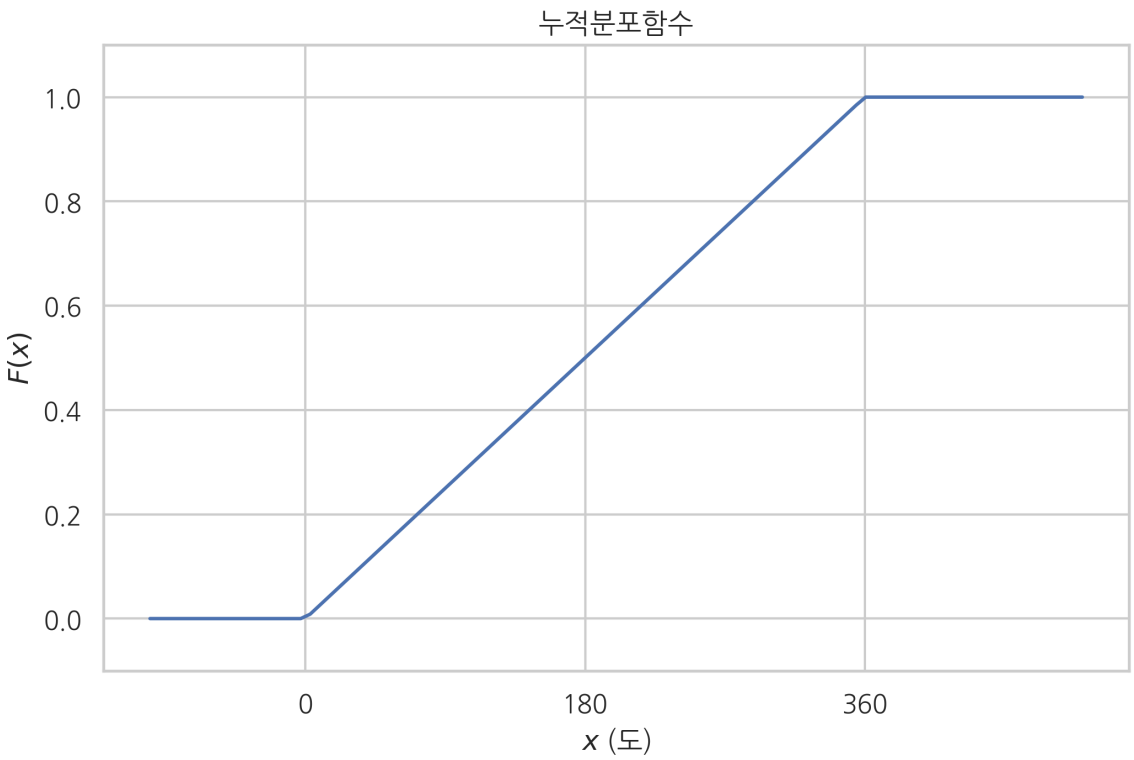
원반의 각도 문제에서 누적분포함수는 다음과 같다. 이 경우에는 각도가 0도부터 360까지이지만 음의 무한대를 시작점으로 해도 상관없다.

$$
\begin{aligned}
F(-10) &= P(\{-\infty^\circ < \theta \leq -10^\circ\}) = 0 \\
F(0) &= P(\{-\infty^\circ < \theta \leq 0^\circ\}) = 0 \\
F(10) &= P(\{-\infty^\circ < \theta \leq 10^\circ\}) = \frac{1}{36} \\
F(20) &= P(\{-\infty^\circ < \theta \leq 20^\circ\}) = \frac{2}{36} \\
&\vdots \\
F(350) &= P(\{-\infty^\circ < \theta \leq 350^\circ\}) = \frac{35}{36} \\
F(360) &= P(\{-\infty^\circ < \theta \leq 360^\circ\}) = 1 \\
F(370) &= P(\{-\infty^\circ < \theta \leq 370^\circ\}) = 1 \\
&\vdots
\end{aligned}
\tag{6.4.28}
$$

이를 NumPy와 matplotlib를 사용하여 그래프로 그리면 다음과 같다.

In [2]:

```
t = np.linspace(-100, 500, 100)
F = t / 360
F[t < 0] = 0
F[t > 360] = 1
plt.plot(t, F)
plt.ylim(-0.1, 1.1)
plt.xticks([0, 180, 360])
plt.title("누적분포함수")
plt.xlabel("$x$ (도)")
plt.ylabel("$F(x)$")
plt.show()
```



연습 문제 6.4.4

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

확률밀도함수

누적분포함수는 1차원 함수라는 편리한 도구를 사용하여 확률분포를 간결하고 정확하게 묘사할 수 있도록 해 주었다.

그러나 누적분포함수가 표현하는 사건이 음수 무한대를 시작점으로 하고 변수 x를 끝점으로 하는 구간이다보니 분포의 형상을 직관적으로 이해하기는 힘든 단점이 있다. 다시 말해서 어떤 확률 변수 값이 더 자주 나오는지에 대한 정보를 알기 힘들다는 점이다.

이를 알기 위해서는 확률 변수가 나올 수 있는 전체 구간 ($-\infty$ ~ ∞)을 아주 작은 폭 dx를 가지는 구간들로 나눈 다음 각 구간의 확률을 살펴보는 것이 편리하다. 만약 x_1 근처에서 폭 dx를 가지는 구간의 확률을 구하면 다음과 같다.

$$P(\{x_1 < x \leq x_1 + dx\}) = F(x_1 + dx) - F(x_1) \tag{6.4.29}$$

이 값은 구간의 길이에 따라 달라지므로 구간 길이 dx를 아주 작게 줄였을 때의 값을 알아야 한다. 단순히 dx를 0으로 줄이면 확률은 당연히 0으로 수렴한다. 우리가 원하는 것은 '같은 구간 길이를 dx를 가진 두 구간이 x_1 위치와 x_2 위치에서 얼마나 다른가'이므로 단위 구간 길이당 확률값으로 비교한다.

그런데 단위 구간 길이당 배정된 확률값

$$\frac{P(\{x_1 < x \leq x_1 + dx\})}{dx} \tag{6.4.30}$$

는 단위 구간이 미세하게 줄어들면 다음 그림에서 보듯이 누적분포함수의 기울기가 된다.

$$\begin{aligned} \lim_{dx \to 0} \frac{P(\{x_1 < x \leq x_1 + dx\})}{dx} &= \lim_{dx \to 0} \frac{F(x_1 + dx) - F(x_1)}{dx} \\ &= x_1\text{에서} F(x)\text{의 기울기} \end{aligned} \tag{6.4.31}$$

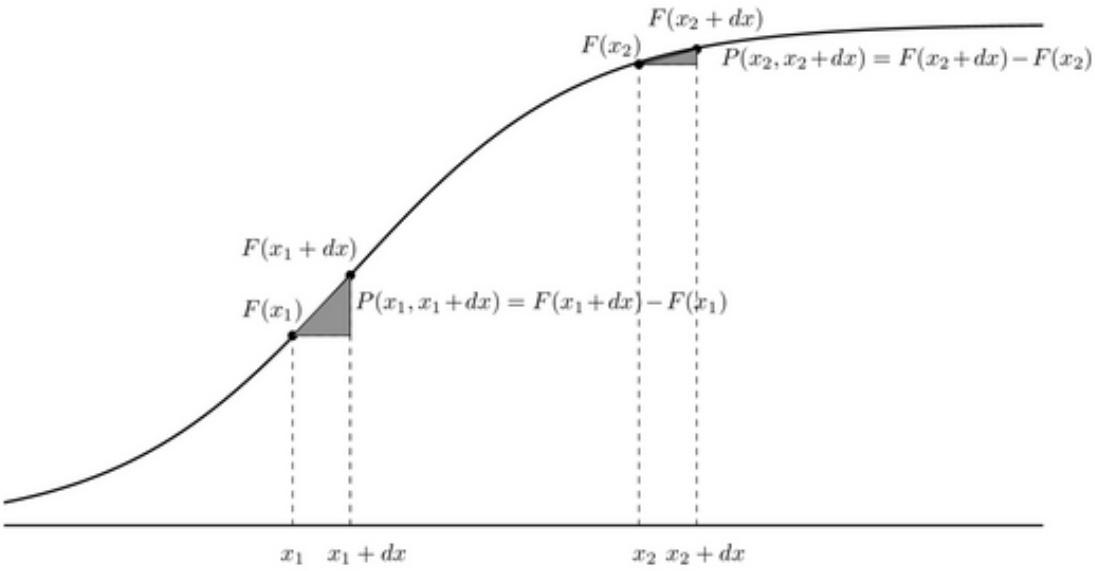


그림 6.4.2 : 누적분포함수에서 구간의 확률을 비교하기

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(probability density function)**라고 한다. 확률질량함수와 마찬가지로 $p(x)$로 표기한다.

$$p(x) = \frac{dF(x)}{dx} \tag{6.4.32}$$

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

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

$$F(x_2) - F(x_1) = \int_{x_1}^{x_2} p(u)du \tag{6.4.33}$$

따라서 누적분포함수와 확률밀도함수의 관계를 적분으로 나타내면 다음과 같다.

$$F(x) = \int_{-\infty}^{x} p(u)du \tag{6.4.34}$$

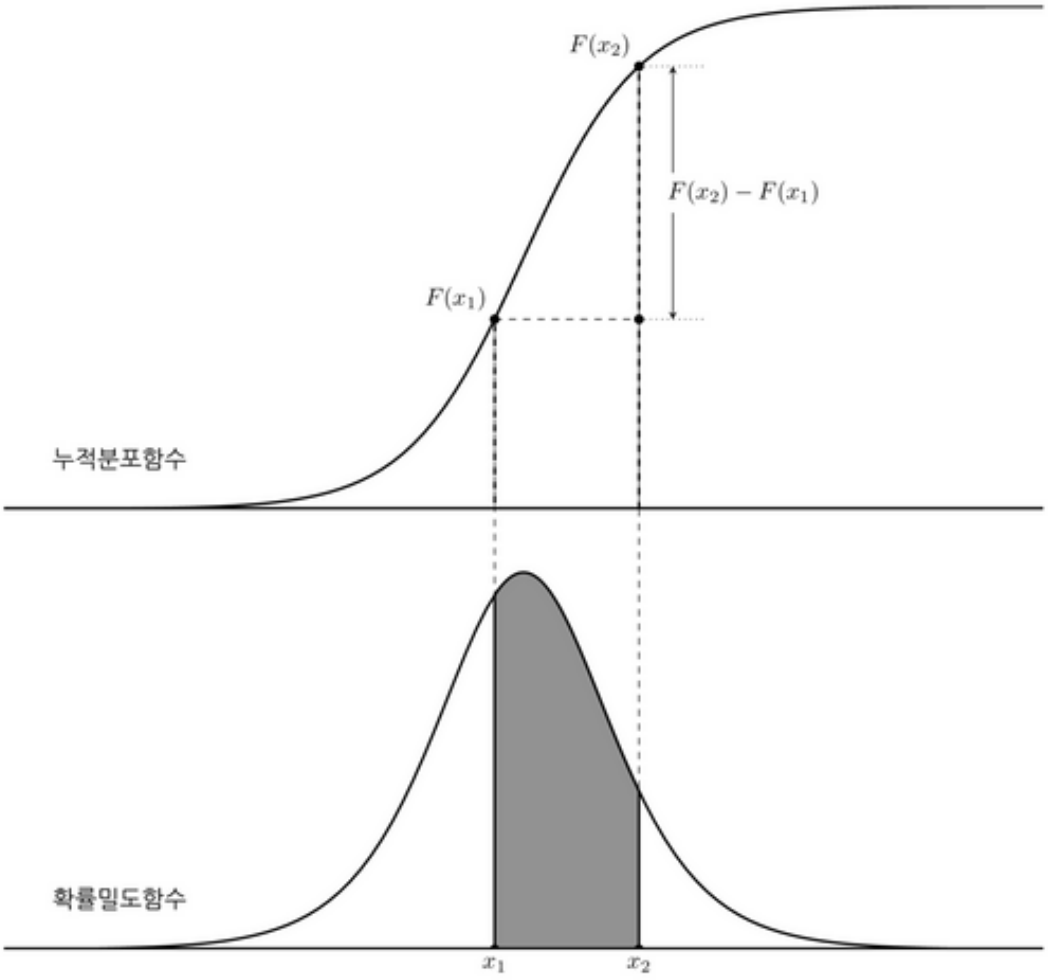


그림 6.4.3 : 확률밀도함수의 면적과 누적분포함수

확률밀도함수는 다음과 같은 특징을 가진다.

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

$$p(x) \geq 0 \tag{6.4.35}$$

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

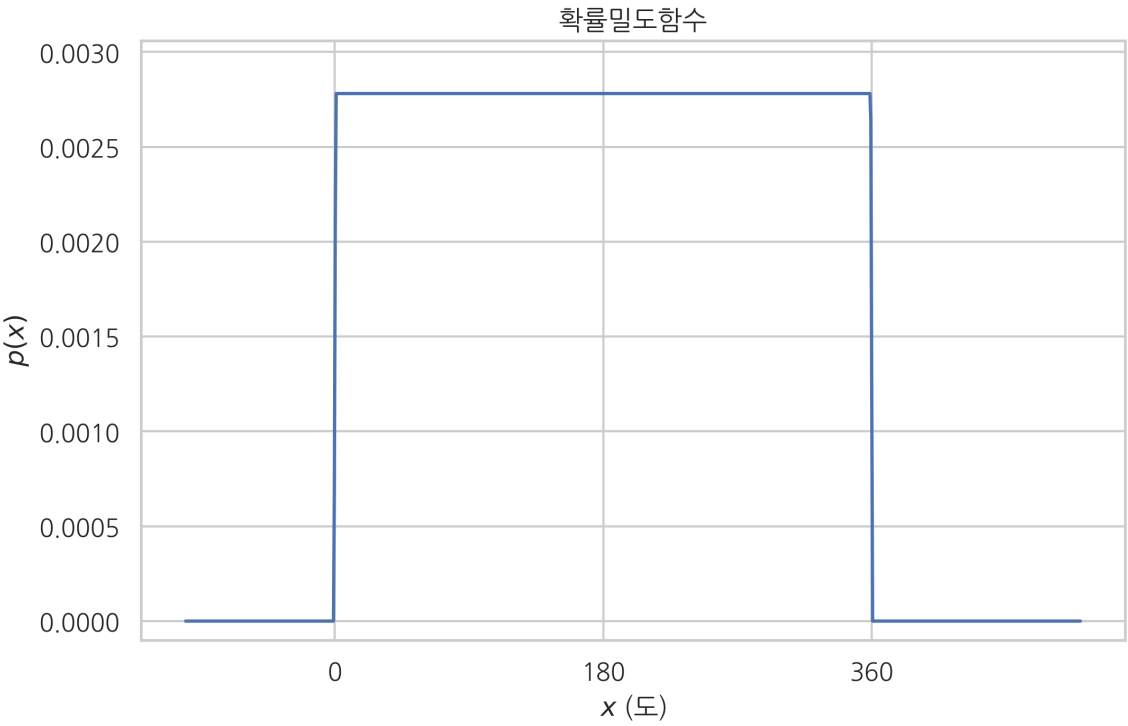
$$\int_{-\infty}^{\infty} p(u)du = 1 \tag{6.4.36}$$

예제

앞의 원반 예제의 확률밀도함수를 구하면 다음과 같다.

In [3]:

```
t = np.linspace(-100, 500, 1000)
F = t / 360
F[t < 0] = 0
F[t > 360] = 1
p = np.gradient(F, 600/1000)  # 수치미분
plt.plot(t, p)
plt.ylim(-0.0001, p.max()*1.1)
plt.xticks([0, 180, 360])
plt.title("확률밀도함수")
plt.xlabel("$x$ (도)")
plt.ylabel("$p(x)$")
plt.show()
```



연습 문제 6.4.5

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