7.2 기댓값과 확률변수의 변환

~~정리 7.2.1~~

확률변수의 기댓값

확률변수의 확률질량함수 ⟶ 확률변수의 이론적 평균값 = 기댓값 (expectation)

- 확률변수 X의 기댓값을 구하는 연산자 (operator) : $E[X]$
- 확률변수 X의 기댓값 : μ_X 또는 μ

* 이산확률변수의 기댓값 = 표본공간의 원소 x_i의 가중평균
(가중치는 x_i가 나올 확률, 즉 pmf $p(x_i)$)

$$\mu_X = E[X] = \sum_{x_i \in \Omega} x_i p(x)$$

x_i : sample(X). 1번씩만 나오는 경우의 수 (O)

예제 7.2.1

공정한 동전이 있고, 이 동전의 앞면이 나오면 1, 뒷면이 나오면 0인 확률변수 X가 있다.
$E[X]$를 구하라. ⟶ $1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}$

(참고: 데이터 공간에서 기댓값에 대응하는 값인 표본평균을 구하는 공식)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

x_i : sample.

기댓값공식과 표본평균 공식에서 x_i의 의미가 다른점이 유의할것
- 기댓값 공식에서 x_i는 표본공간의 모든 원소를 뜻하지만,
표본평균 공식에서 x_i는 선택된 (sampled, realized) 표본만을 뜻한다.

NO.

DATE.

연습문제 7.2.2

기댓값을 구하는 공식에서는 확률을 가중치로 곱한다. 그런데 sample 표본평균을 구하는 공식에서는 확률을 가중치가 없었는가?

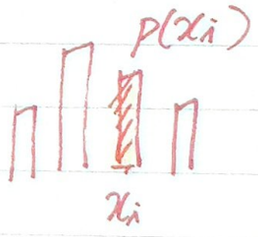
* 연속확률변수의 기댓값은 확률밀도함수 $p(x)$를 가중치로 하여 모든 가능한 표본 x를 적분한 값이다.

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x\,p(x)\,dx$$

기댓값은 여러 가능한 x값을 확률 (또는 확률밀도) 값에 따라 가중합을 한 것으로 각각 확률(또는 확률밀도)이 몰린 x값 근처의 값이 된다. 즉, 확률 (또는 확률밀도)가 모여있는 곳의 위치를 나타낸다.

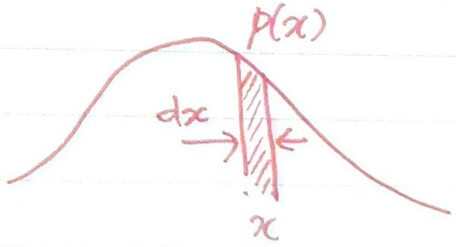
① discrete X

$$E[X] = \sum_{x_i \in \Omega} x_i \underbrace{p(x_i)}_{\text{pmf}}$$



② continuous X

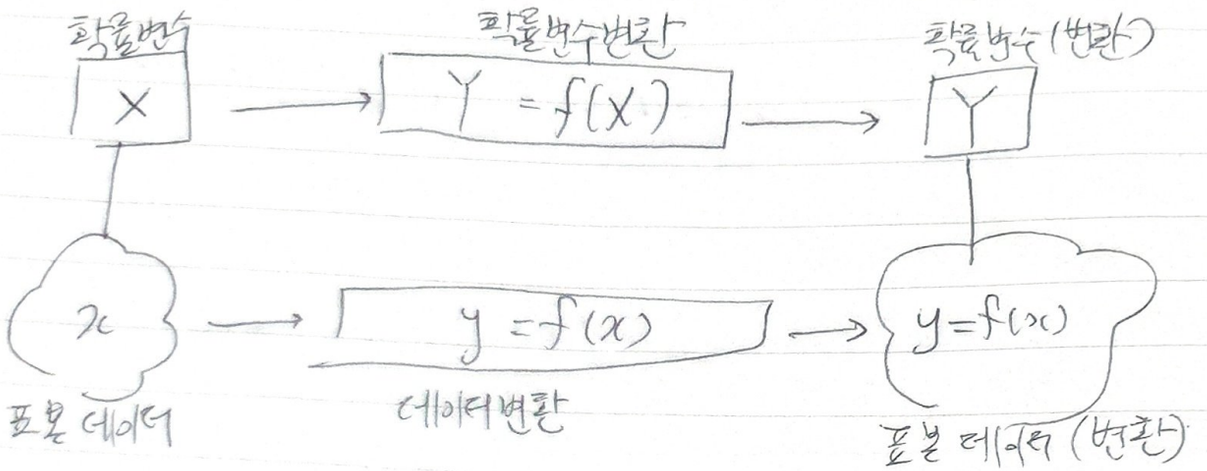
$$E[X] = \int_{-\infty}^{\infty} x \underbrace{p(x)}_{\text{pdf}} dx$$

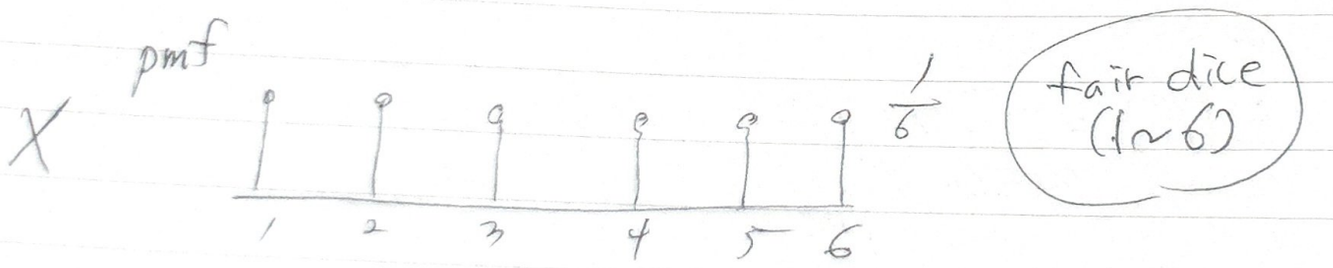


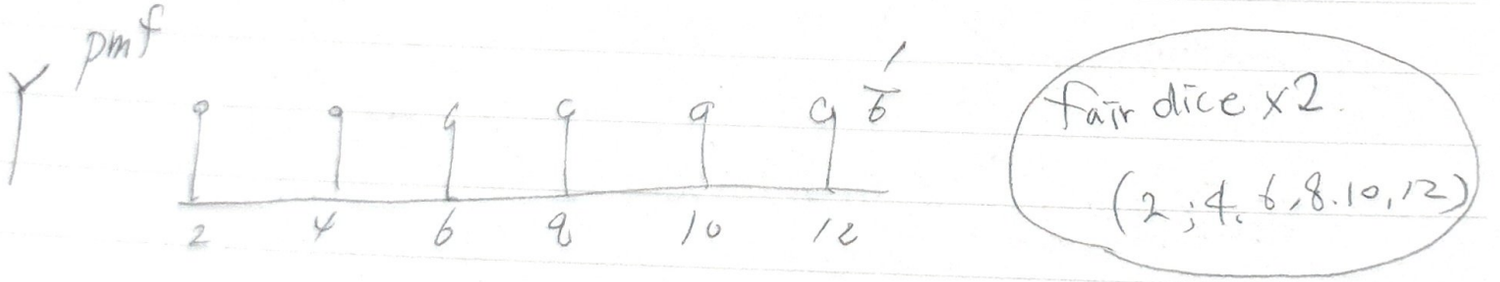
확률변수의 변환

우리가 얻은 데이터의 값을 어떤 함수 f에 넣어서 변환시킨다고 가정하자. 그러면 새로운 데이터 집합이 생긴다.

$$\{x_1, x_2, \cdots, x_N\} \longrightarrow \{f(x_1), f(x_2), \cdots f(x_N)\}$$







NO.

DATE.

확률변수 X에서 표본을 N번 뽑아서 그 값을 더하는 경우에는, 다음처럼 원래 확률변수의 복사본 $X_1, X_2, \cdots, X_N$을 만든 다음 이 복사본 확률변수의 표본값을 더한 형태로 변환식을 써야 한다.

$$Y = X_1 + X_2 + \cdots + X_N$$

이렇게 복사본을 만들어 첨자를 붙이는 이유는 X_1과 X_2가 같은 확률분포를 가지는 확률변수지만 표본값이 다르기 때문이다.

만약 다음과 같이 쓰면,

$$Y = X + X + \cdots + X$$

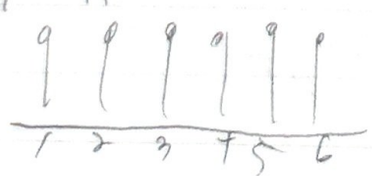
이 Y는 다음처럼 전혀 다른 확률변수를 가리킨다.

$$Y = N \cdot X \neq (X_1 + X_2 + \cdots + X_N)$$

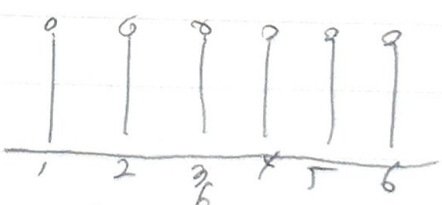
연습문제 7.2.5

확률변수 X_1과 X_2는 각각 주사위를 던져 나오는 수를 나타내는 확률변수다. 그리고 Y는 두 주사위를 동시에 던져 나오는 수의 합을 나타내는 확률변수다. 확률변수 X_1, X_2, Y의 확률질량함수의 그래프를 각각 그려라.

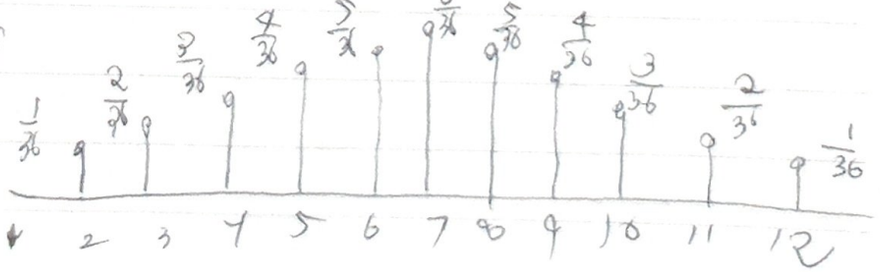
X_1 pmf



X_2 pmf



$Y = X_1 + X_2$ pmf



기댓값의 성질

상수 C에 대해

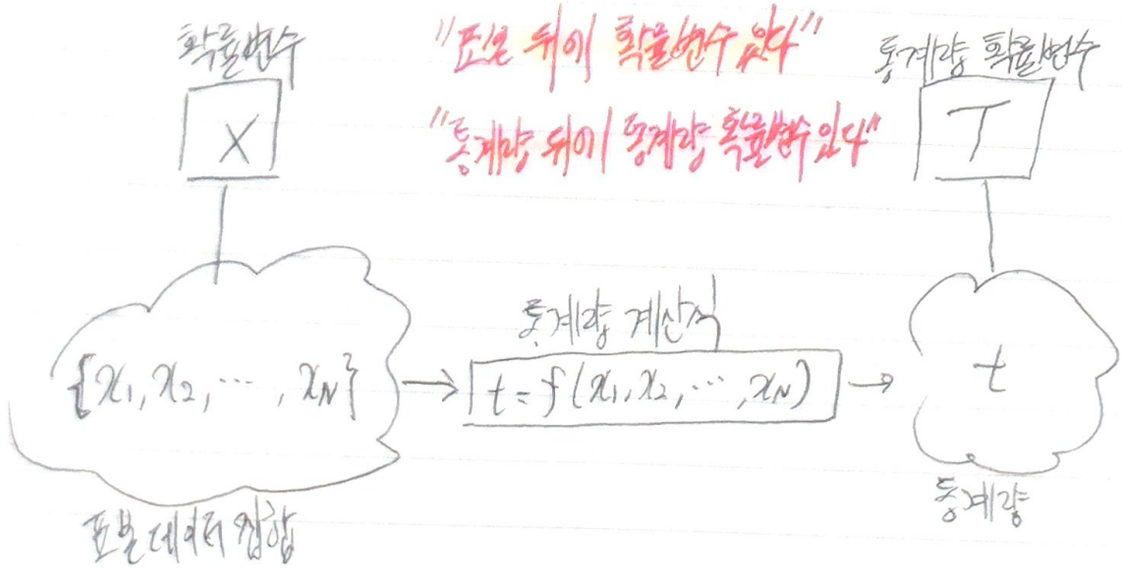
$$E[C] = C$$

《선형성》

$$E[CX] = CE[X]$$
$$E[X+Y] = E[X]+E[Y]$$
$$E[C_1X + C_2Y] = C_1E[X] + C_2E[Y]$$

통계량 Statistics

확률변수 X로부터 데이터 집합 $\{x_1, x_2, \cdots, x_N\}$을 얻었다고 하자.
이 데이터 집합 (sample)의 모든 값을 정해진 어떤 공식에 넣어서 하나의 숫자를 구한 것을 통계량 (Statistics)이라고 한다.
예를 들어 표본의 합, 표본평균, 표본중앙값, 표본분산 등은 모두 통계량이다.
통계량도 확률변수의 변환에 포함된다.



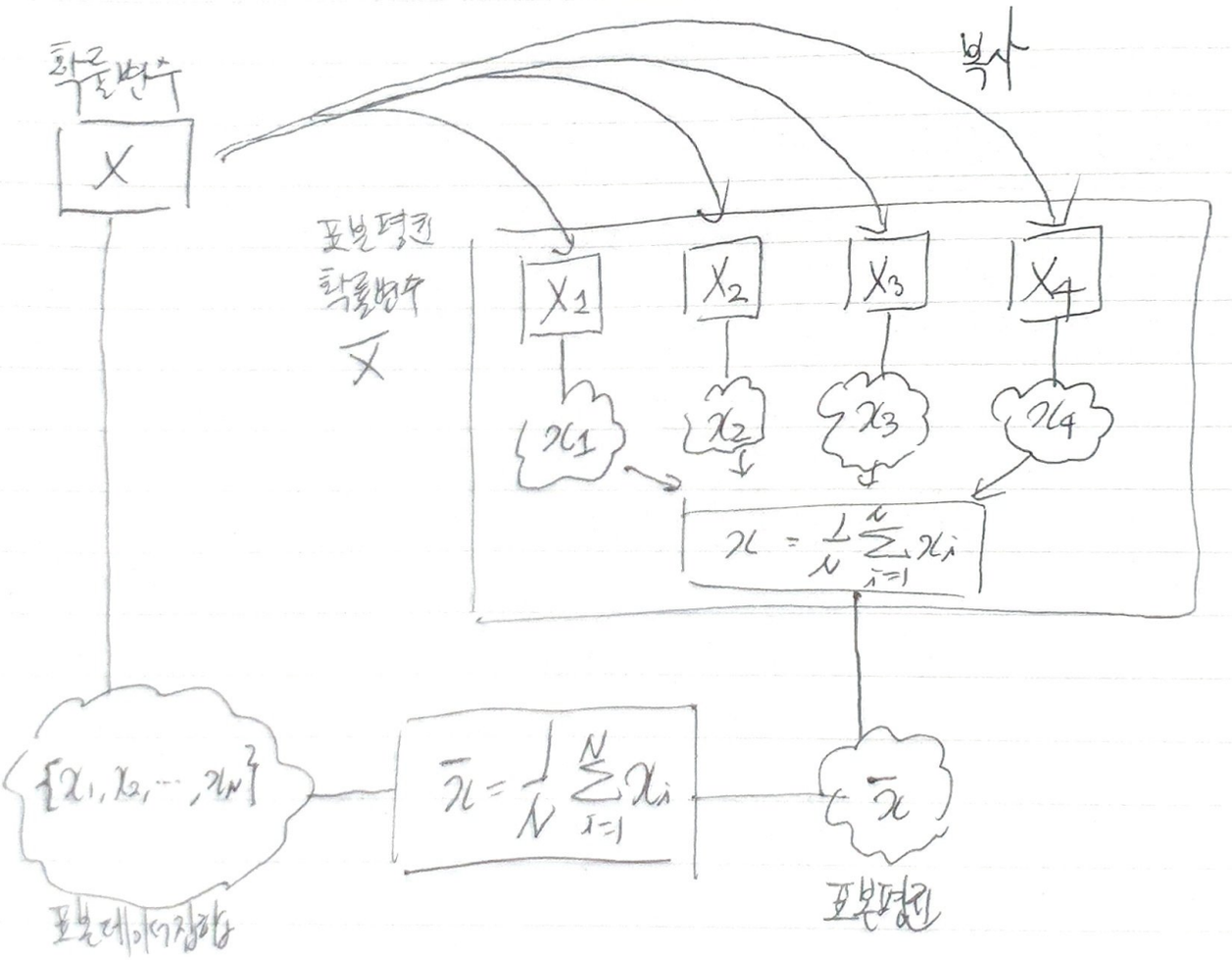
NO.

DATE

표본평균 확률변수

확률변수로부터 N개의 표본을 만들어 이 표본집합의 표본평균을 구하면 이렇게 구한 표본평균 값도 확률변수가 된다. 표본평균 확률변수는 일래의 확률변수 위에 윗줄(bar)을 추가하여 $\bar{X}$와 같이 표기한다.

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$



(연습문제 7.2.6)

표본평균 $\bar{x}$의 값은 확률적인 데이터이고, 이를 생성하는 확률변수 $\bar{X}$는 다음과 같이 정의할 수 있었다. 그렇다면 (편향) 샘플분산 s^2의 값은 확률적인 데이터인가? 만약 그렇다면 이를 생성하는 확률변수 S^2은 어떻게 정의해야 하는가?

$$\bar{x} = \frac{1}{N}\sum x_i$$

$$s^2 = \frac{1}{N}\sum \left(x_i - \frac{1}{N}\sum x_i\right)^2$$

기댓값과 표본평균의 관계

표본평균도 확률변수이고 기댓값이 존재한다. 표본평균의 기댓값은 원래의 확률변수의 기댓값과 같다는 것을 다음처럼 증명할 수 있다.

$$E[\bar{X}] = E[X]$$

표본평균 분포에서 확률에 따른 값 = 원래 확률변수 분포에서 확률에 따른 값

(증명)

$$E[\bar{X}] = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[X_i]$$

X_i는 X의 copy이므로 기댓값은 같다.

$$= \frac{1}{N}\sum_{i=1}^{N} E[X]$$

$$= \frac{1}{N}\cdot N \cdot E[X]$$

$$= E[X]$$

표본평균은 확률변수의 기댓값과 같은 값이다.

$E[\text{fair dice}] = 3.5 \longrightarrow E[\overline{\text{fair dice}}] \approx 3.5$

NO.

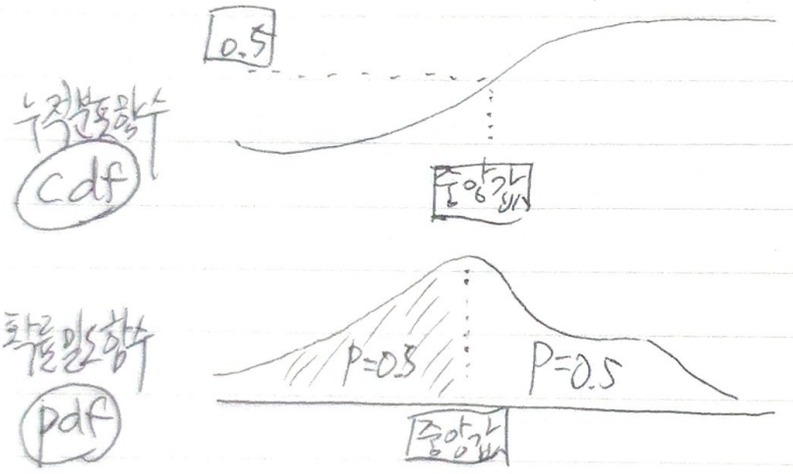
DATE

중앙값

확률변수의 중앙값(median)은 중앙값보다 큰 값이 나올 확률과 작은 값이 나올 확률이 0.5로 같은 값을 뜻한다. 따라서 다음과 같이 누적확률분포 $F(x)$ 에서 중앙값을 계산할 수 있다.

$$0.5 = F(\text{중앙값})$$

$$\text{중앙값} = F^{-1}(0.5)$$



최빈값

most frequent value

- 이산확률분포에서 — "가장 확률값이 큰 수" = 최빈값(mode)
- 연속확률분포에서 — "확률밀도함수 pdf $p(x)$의 값이 가장 큰 확률변수의 값" (확률밀도함수의 최댓값의 위치)

$$\text{최빈값} = \arg\max_x p(x)$$