확률 분포

윤 정 훈

들어가며

- 동전을 3회 독립적으로 반복하여 던지는 실험

  가능한 모든 사건의 집합인 표본공간 S는 $\{HHH, HHT, HTT, HTH, HTT, TTH, TTT, THH\}$이다.

- 주사위를 1회 던지는 실험

  가능한 모든 사건의 집합인 표본공간 S는 $\{1,2,3,4,5,6\}$이다.

✓ 이처럼 표본공간 S는 수치적으로 또는 비수치적으로 표현될 수 있다. 많은 경우 실험결과로부터 계산될 수 있는 어떤 **수치적인 양**이 관심의 대상이 된다. 예를 들어, 동전을 3회 독립적으로 반복하여 던지는 첫번째 실험에서 표본공간은 HHH , HHT , HTH 등으로 이루어지지만, 이것을 동전의 앞면이 나오는 횟수인 3 , 2 , 2 등으로 생각할 수 있다. 즉, $\{HHH, HHT, HTT, HTH, HTT, TTH, TTT, THH\}$를 앞면이 나온 횟수의 수치적 집합인 $\{3,2,1,2,1,1,0,2\}$로 표현할 수 있는 것이다.

✓ 표본공간의 사건들을 수치적 양으로 나타내고, 수치적 양에 대한 수리적 모형을 생각해보자.

확률변수

표본공간 S에 정의된 실함수(real-valued function)이다. 즉 $s \in S$, $r \in R$, $X(s) = r$

ex) 동전을 3회 반복하여 던졌을 경우,
앞면이 나오는 횟수에 대한 확률변수 X

s	$X(s)$ or X
HHH	3
HHT	2
HTH	2
THH	2
HTT	1
THT	1
TTH	1
TTT	0

확률변수의 예제

ex) 동전을 3회 반복하여 던졌을 경우,
앞면이 나오는 횟수에 대한 확률변수 X

s	$X(s)$ or X
HHH	3
HHT	2
HTH	2
THH	2
HTT	1
THT	1
TTH	1
TTT	0

동전의 앞면이 나올 확률 : p
동전의 뒷면이 나올 확률 : $1-p$

확률변수 X에 대응되는 확률분포

- ✓ $P(X=0)=P(\{TTT\})=(1-p)^3$
- ✓ $P(X=1)=P(\{HTT,THT,TTH\})=3p(1-p)^2$
- ✓ $P(X=2)=P(\{HHT,HTH,THH\})=3p^2(1-p)^1$
- ✓ $P(X=3)=P(\{HHH\})=p^3$

엄밀히! S에서의 P와 X에서의 P는 달라야 한다.

확률변수의 종류

이산형

X가 가질 수 있는 수의 범위가 유한(finite)이거나, 가산무한(countably infinite)이면 이 확률변수 X는 이산형확률변수(discrete random variable)라고 한다.

ex) 주사위를 던졌을 때 나오는 눈을 확률변수로 표현 $X = \{1,2,3,4,5,6\}$

연속형

X가 가질 수 있는 수의 범위를 셀 수 없거나, 불가산무한(uncountably infinite)이면 이 확률변수 X는 연속형 확률변수(continuous random variable)라고 한다.

ex) 분당지역 중학생들의 키를 확률변수로 표현

확률질량함수 (이산형 확률분포)

확률질량함수(probability mass function) f는, 이산형 확률변수 X를 정의역으로 하며,

- 모든 실수 x에 대하여, $f(x) \geq 0$이다.
- 확률변수 X가 가질 수 있는 유한 또는 가산무한개의 값 $x_1, \ldots$ 에 대하여 $f(x_i) \geq 0$이며 $\sum_i f(x_i) = 1$이다.

이산형 확률분포는 이산형 확률변수 X가 가질 수 있는 가능한 값 $x_1, x_2, \ldots, x_n$과 이에 대응하는

확률 $P(X = x_i),\ \ i = 1, \ldots, n$ 을 함수로 나타내면 되고, 이 때의 함수가 확률질량함수 $f(x)$이다.

확률질량함수 $f(x)$는 $P(X = x) = f(x)$를 만족.

즉, discrete 한 성격을 가진 확률변수에 대하여, 각각의 확률값이 정의된다.

확률질량함수 (이산형 확률분포) 예제

○ 한 개의 동전을 3번 던지는 실험에서 앞면이 나오는 횟수를 X라 하자.

① 확률변수 X의 확률분포를 구하여라.

② 앞면이 두 번 이상 나올 확률을 구하여라.

확률밀도함수 (연속형 확률분포)

확률밀도함수(probability density function) f는, 연속형 확률변수 X를 정의역으로 하며,

- 모든 실수 x에 대하여, $f(x) \geq 0$이다.
- 확률변수 X가 가질 수 있는 모든 실수 x_i에 대하여 $f(x_i) \geq 0$이며 $\int_{-\infty}^{+\infty} f(x)dx = 1$이다.

연속형 확률변수 X의 확률밀도함수 $f(x)$는 $P(X = x) = 0$ 이다.

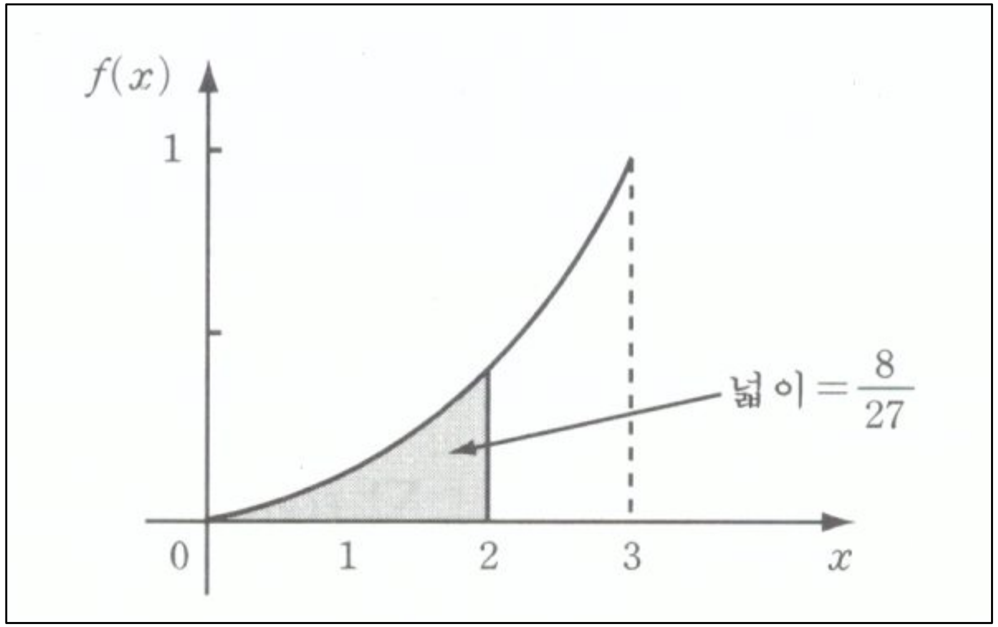
$\int_a^b f(x)dx = P(a \leq x \leq b)$와 같이 구간에 대한 확률값을 정의할 수 있다.

확률밀도함수 (연속형 확률분포) 예제 1

- 구간 (0, 3)에 정의된 함수 $f(x) = \frac{x^2}{9}$는 확률밀도함수이다.
  - ✓ $f(x) \geq 0$
  - ✓ $\int_0^3 f(x)\,dx = 1$

- 연속확률변수 X의 확률밀도함수가 $f(x)$이면, $0 < X < 2$일 확률 :

  $P(0 < X < 2) = \int_0^2 \frac{x^2}{9} dx = \frac{8}{27}$



- 연속확률변수가 어떤 구간에 들어갈 확률은 그 구간에서 확률밀도함수의 적분

확률밀도함수 (연속형 확률분포) 예제 2

○ 어떤 전기시스템에서의 발생하는 전압 X의 확률밀도함수가 다음과 같다고 하자.

$$f(x) = \begin{cases} \dfrac{3}{8}x^2, & 0 \le x \le 2 \\ 0, & \text{기타} \end{cases}$$

① 확률밀도함수 $f(x)$의 그래프를 그려라.

② 확률 $P(1 \le x \le 2)$를 구하여라.

누적분포함수

확률변수 X에 대한 누적(확률)분포함수(cumulative distribution function) $F(X)$는

$$F(X) = P(X \leq x)$$

로 정의된다.

- ✓ 누적(확률)분포함수는 확률변수 X가 주어진 점 x 이하의 모든 값을 가질 확률을 누적
- ✓ 구간에 대한 확률값 : $P(a < X \leq b) = F(b) - F(a)$
- ✓ 연속형확률변수 X의 확률밀도함수가 $f(x)$이고 확률분포함수가 $F(x)$이면, $f(x) = \frac{d}{dx}F(x)$

누적분포함수 예제

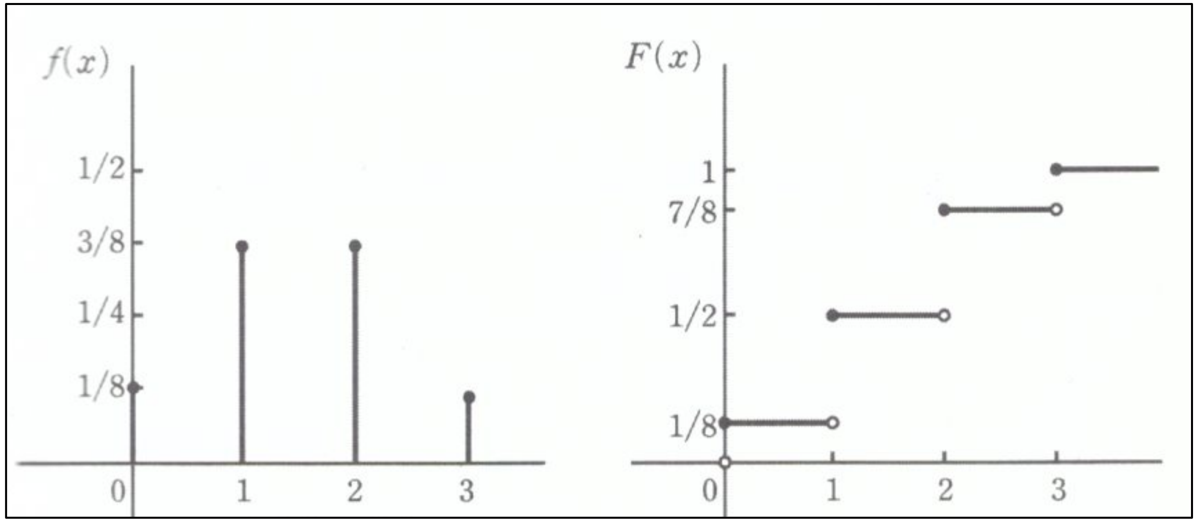
- 동전을 3회 독립반복하여 던졌을 경우 관심있는 변수 X(=앞면의 수)
  - ✓ 확률질량함수

$$P(X=0)=P(\{TTT\})=(1-p)^3$$
$$P(X=1)=P(\{HTT,THT,TTH\})=3p(1-p)^2$$
$$P(X=2)=P(\{HHT,HTH,THH\})=3p^2(1-p)^1$$
$$P(X=3)=P(\{HHH\})=p^3$$

▷ $f(x)=\begin{cases}\frac{1}{8} & x=0\\ \frac{3}{8} & x=1 \text{ or } 2\\ \frac{1}{8} & x=3\end{cases}$

  - ✓ 확률분포함수(누적분포함수)

▷ $F(x)=\begin{cases}0 & x<0\\ \frac{1}{8} & 0\le x<1\\ \frac{1}{2} & 1\le x<2\\ \frac{7}{8} & 2\le x<3\\ 1 & x\ge 3\end{cases}$



확률질량함수 확률분포함수

결합 확률밀도함수(2변수)

여러 개의 확률변수들을 한꺼번에 고려하는 경우, 결합분포 (joint distribution) 이론이 필요

두 확률변수 X와 Y의 결합 확률밀도함수 $f_{X,Y}(x,y)$는

- X, Y가 이산형인 경우

$$f_{X,Y}(x,y) = P(X = x, Y = y)$$

로 정의되며 (결합 확률질량함수라고도 함),

- X, Y가 연속형인 경우 이차평면상의 임의의 영역 A에 대하여,

$$P[(X,Y) \in A] = \iint_A f_{X,Y}(x,y)dxdy$$

를 만족하는 $f_{X,Y}(x,y)$로 정의된다.

결합 확률밀도함수 예제

- (이산형) 4개의 빨간 공과 3개의 하얀 공과 2개의 검은 공이 들어 있는 주머니에서 3개의 공을 임의로 꺼낸다고 하자.
  - ✓ X =하얀 공의 수
  - ✓ Y =검은 공의 수
  - ✓ 두 변수 X, Y의 결합 확률밀도함수$(x = 0,1,2,3\ ;\ \ y = 0,1,2\ ;\ \ 0 \leq x + y \leq 3)$:

$$f_{X,Y}(x, y) = \frac{\binom{3}{x}\binom{2}{y}\binom{4}{3-x-y}}{\binom{9}{3}}$$

결합 확률밀도함수(일반형)

k변수로 일반화하자.

확률변수 $X_1, X_2, \dots, X_k$의 결합 확률밀도함수는

✓ 모든 $(x_1, x_2, \dots, x_k) \in R^k$에 대하여, $f(x_1, x_2, \dots, x_k) \geq 0$이다.

- $X_1, X_2, \dots, X_k$가 이산형인 경우
  - ✓ $\sum \dots \sum_{\forall (x_1, \dots, x_k)} f(x_1, x_2, \dots, x_k) = 1$
- $X_1, X_2, \dots, X_k$가 연속형인 경우
  - ✓ $\int \dots \int_{R^k} f(x_1, \dots, x_k)\, dx_1 \dots dx_k = 1$

k차원 공간상의 임의의 영역 $A \subset R^k$에 대하여 $(x_1, x_2, \dots, x_k) \in A$일 확률

$$P[(x_1, \dots, x_k) \in A] = \begin{cases} \displaystyle\sum \cdots \sum_{(x_1, \dots, x_k) \in A} f(x_1, \dots, x_k) & \text{이산형} \\ \displaystyle\int \cdots \int_A f(x_1, \dots, x_k) dx_1 \dots dx_k & \text{연속형} \end{cases}$$

결합 누적분포함수(일반형)

확률변수 $X_1, X_2, \dots, X_k$의 결합 확률(누적)분포함수는

- ✓ $F(x_1, x_2, \dots, x_k) = P(X_1 \le x_1, X_2 \le x_2, \dots, X_k \le x_k)$

- ● $X_1, X_2, \dots, X_k$ 가 이산형인 경우
  - ✓ $F(x_1, x_2, \dots, x_k) = \sum \dots \sum_{\forall t_i \le x_i} f(t_1, \dots, t_k)$
- ● $X_1, X_2, \dots, X_k$ 가 연속형인 경우
  - ✓ $F(x_1, x_2, \dots, x_k) = \int_{-\infty}^{x_k} \dots \int_{-\infty}^{x_1} f(t_1, \dots, t_k) dt_1 \dots dt_k$

로 정의된다.

결합 확률밀도함수는 결합 누적분포함수를 편미분하여 구할 수 있다.

$$f(x_1, x_2, \dots, x_k) = \frac{\partial^k}{\partial x_1 \dots \partial x_k} F(x_1, x_2, \dots, x_k)$$

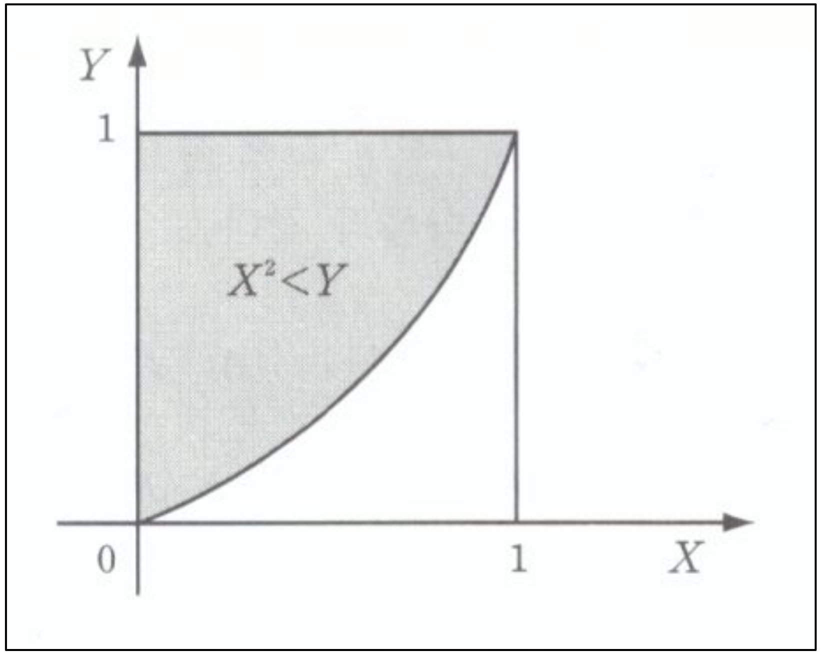
결합 누적분포함수 예제

- 두 변수 X, Y의 결합 확률분포함수가 $F_{X,y}(x, y) = xy$ $(0 < x < 1$, $0 < y < 1)$ 로 주어졌을 때, X, Y의 결합 확률밀도함수 $f_{X,Y}(x, y)$를 구하고 $P(X^2 < Y)$를 계산하라.

✓ $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} xy = 1$

✓ $P(X^2 < Y) = \int_0^1 \int_0^{\sqrt{y}} 1\, dx dy = \frac{2}{3}$

확률을 계산하기 위하여 적분되는 영역



주변 확률밀도함수(2변수)

두 확률변수 X, Y의 결합 확률밀도함수가 $f_{X,Y}(x, y)$로 주어졌을 때, 두 변수 X, Y 각각의 확률밀도함수 $f_X(x)$와 $f_Y(y)$를 주변 확률밀도함수라고 한다.

(marginal probability density function)

X와 Y 각각의 주변 확률밀도함수 $f_X(x)$와 $f_Y(y)$는

- 이산형인 경우, $f_X(x) = \sum_{\forall y} f_{X,Y}(x, y)$, $f_Y(y) = \sum_{\forall x} f_{X,Y}(x, y)$ (주변 확률질량함수라고도 함)
- 연속형의 경우, $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$, $f_Y(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$

주변 확률밀도함수(일반형)

k변수로 일반화하자.

확률변수 $X_1, \dots, X_k$의 결합 확률밀도함수가 $f(x_1, \dots, x_k)$일 때, $X_i(1 \le i \le k)$의 주변 확률밀도함수는,

- 이산형의 경우 :
  - ✓ $\sum \cdots \sum_{x_i}$를 제외한 모든 값 $f(x_1, \dots, x_k)$

- 연속형의 경우 :
  - ✓ $\int \dots \int f(x_1, \dots, x_k)\, dx_1 \dots dx_{i-1} dx_{i+1} dx_k$

로 정의된다.

결합/주변 확률밀도함수 예제

(이산형) 어느 지역에 거주하는 가정 중에서 15%는 자녀가 없고, 20%는 1명, 35%는 2명 30%는 3명의 자녀가 있으며, 더구나 각 가정에서 자녀가 아들과 딸일 가능성은 독립적으로 동일하다고 가정한다. 만일 이 지역에서 한 가정이 무작위로 선택되었다면, 이 가정에서 아들의 수 B와 딸의 수 G는 다음과 같은 확률값을 가진다.

- P(B=0, G=0) = P(자녀가 없음) = 0.15
- P(B=0, G=1) = P(자녀가 1명으로 딸인 경우) = P(자녀가 1명)P(딸이 한명|자녀가 1명) = $0.20\times(\frac{1}{2})$
- P(B=0, G=2) = P(자녀가 2명으로 딸인 경우) = P(자녀가 2명)P(딸이 한명|자녀가 2명) = $0.35\times\left(\frac{1}{2}\right)^2$

결합/주변 확률밀도함수 예제

이산형이므로, 결합/주변 확률질량함수를 아래와 같이 나타낼 수 있다.

결합 확률질량함수 : $f_{B,G}(i,j) = P(B=i, G=j)$

i \ j	0	1	2	3	행의 합 = $P(B=i)$
0	0.15	0.10	0.0875	0.0375	0.3750
1	0.10	0.175	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2000
3	0.0375	0	0	0	0.0375
열의 합 = $P(G=j)$	0.375	0.3875	0.2000	0.0375	

$P_B(i)$

$P_G(j)$

주변 확률질량함수

독립확률변수

두 확률변수 X와 Y는 임의의 실구간 A와 B에 대하여

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

가 성립할 때, 서로 독립(independent)이라고 한다.

이것을 확률 밀도함수를 이용하여 나타내면,

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

독립확률변수 예제

- 두 변수 X, Y는 서로 독립인가?

  ✓ 두 변수 X, Y의 결합 확률밀도함수

  $$f_{X,Y}(x,y) = \frac{\exp(-4)\cdot 2^{x+y}}{x!y!}\ ,\ x, y = 0,1,2,\ldots$$

  ✓ X의 주변 확률밀도함수

  $$f_X(x) = \sum_y f_{X,Y}(x,y) = \frac{\exp(-2)\cdot 2^{x}}{x!}\ ,\ x = 0,1,2,\ldots$$

  ✓ Y의 주변 확률밀도함수

  $$f_Y(y) = \sum_x f_{X,Y}(x,y) = \frac{\exp(-2)\cdot 2^{y}}{y!}\ ,\ y = 0,1,2,\ldots$$

  $\underline{f_{X,Y}(x,y) = f_X(x)\cdot f_Y(y)}$이므로 $\underline{X\text{와 } Y\text{는 독립}}$이다.

확률분포의 특성을 요약

확률밀도함수나 누적분포함수는 확률변수의 전체적인 성격을 설명한다. 때로 우리는 몇 개의 수치로 확률분포의 성질을 요약하고자 하기도 한다. 이러한 확률분포의 성질을 요약하는 수치들은 다음과 같다.

- 변수의 기댓값
- 변수의 분산
- 변수의 표준편차
- 변수의 중간값
- 변수의 최빈값

기댓값

확률변수 X의 확률밀도함수가 $f(x)$일 때, X의 기대값은 다음과 같이 정의된다.

- X가 이산형인 경우
  - ✓ $E(X) = \sum_{\forall x_i} x_i f(x_i)$
- X가 연속형인 경우
  - ✓ $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$

$Y = g(X)$인 경우, $g(X)$의 기대값

- 이산형 변수 : $E(X) = \sum_{\forall x_i} g(x_i) f_X(x_i)$
- 연속형 변수 : $E(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$

기대값 예제 1

- 동전을 2번 던지는 실험에서 앞면이 나오는 횟수 X라고 하자.

① 횟수 X의 확률분포는?

② 이 때의 기대값은?

기대값 예제 2

- 연속형 확률변수 X의
  - ✓ 확률밀도함수 :

    $f(x) = \frac{x^2}{9}$ (단, $0 < x < 3$)
  - ✓ X의 기대값 :

    $$E(X) = \int_0^3 xf(x)\,dx = \int_0^3 \frac{x^3}{9}dx = \frac{9}{4}$$

- 연속형 확률변수 X의
  - ✓ 확률밀도함수

    $f(x) = xe^{-x}$
  - ✓ X의 기대값 :

    $$E(X) = \int_0^\infty xf(x)\,dx = \int_0^\infty x^2e^{-x}\,dx = 2$$

기대값 예제 3

- 연속형 확률변수 X에 대하여

  ✓ 확률밀도함수 :

  $$f_X(x) = xe^{-x} \text{ , } (x > 0)$$

  ✓ $g(X) = X^2 + 5$의 기대값 :

  $$E(X^2 + 5) = \int_0^\infty (x^2 + 5)\, f_X(x)dx$$

  $$= \int_0^\infty x^3 e^{-x}\, dx + \int_0^\infty 5xe^{-x}\, dx$$

  $$= 11$$

기댓값의 성질

- 확률변수 X의 함수에 대한 기댓값은, 상수 a, b, c 에 대하여 다음을 만족한다.
  - ✓ $E(c) = c$
  - ✓ $E(aX + b) = aE(X) + b$

- 두 확률변수 X와 Y가 서로 독립이면, 다음이 성립한다.
  - ✓ $E(XY) = E(X) \cdot E(Y)$

분산 / 표준편차

- 분산 (variance)
  - ✓ 확률변수 X의 <u>변동</u> 또는 <u>흩어짐</u>을 나타내는 측도로서, X가 평균 $E(X)$로부터 흩어져 있는(또는 밀집된) 정도
  - ✓ $\boldsymbol{Var(X) = E[X - E(X)]^2}$ ⟶ - (이산형) $\Sigma(x_i - \mu)^2 f(x_i)$

    - (연속형) $\int (x - \mu)^2 f(x) dx$
- 표준편차 (standard deviation)
  - ✓ 분산의 제곱근
  - ✓ $\boldsymbol{\sigma_X = \sqrt{Var(X)}}$

분산 / 표준편차 예제1

- 동전을 3번 던졌을 때 앞면이 나오는 횟수를 X라고 하자.

① 횟수 X의 분포를 구하여라.

② 횟수 X의 평균을 구하여라.

③ 횟수 X의 분산 및 표준편차를 구하여라.

분산 / 표준편차의 성질

○ 실제 문제에 있어서 분산을 이용하여 변수 X의 흩어짐을 재는 것은 단위의 문제가 발생한다. 따라서 **분산의 제곱근**으로 단위가 통일된 표준편차를 사용하여 변수의 흩어짐을 측정한다.

○ 확률변수 X를 그의 기대값과 표준편차를 사용하여 $Z = \frac{X-E(X)}{\sigma_X}$로 변환을 한다. 이것을 **변수의 표준화**(standardization)이라고 한다.

○ $Var(X) = E(X^2) - [E(X)]^2$

분산 예제

- 연속형 확률변수 X에 대하여,

  ✓ X의 확률밀도함수 :

  $f(x) = \frac{4(1-x^3)}{3}$, $(0 < x < 1)$

  ✓ X의 기대값

  $$E(X) = \int_0^1 x \frac{4(1-x^3)}{3} dx = \frac{2}{5}$$

  ✓ X^2의 기대값

  $$E(X^2) = \int_0^1 x^2 \frac{4(1-x^3)}{3} dx = \frac{2}{9}$$

  ✓ X의 분산

  $$Var(X) = E(X^2) - [E(X)]^2 = \frac{2}{9} - \frac{4}{25} = \frac{14}{225}$$

분산 / 표준편차의 성질

- 확률변수 X, Y에 대하여, $Y = aX + b$이면 다음이 성립한다.
  - ✓ $Var(Y) = a^2 Var(X)$

- 확률변수 $X_1, X_2, \dots, X_n$이 서로 독립이면 다음이 성립한다.
  - ✓ $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$

공분산

- 공분산(covariance)은 X와 Y가 선형적으로 함께 움직이는 정도를 표현
- $Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

공분산은 X와 Y의 선형관계에 대한 측도이다.

X가 증가할 때 Y가 증가하는 경향이 있으면 양의 값을,

X가 증가할 때 Y가 감소하는 경향이 있으면 음의 값을 가진다.

측정단위의 영향을 받는 단점이 있다.

체비세프 부등식

확률변수 X의 평균이 μ이고 분산이 $\sigma^2 < \infty$이면, 임의의 $k > 0$에 대하여, 다음이 성립

$$P[|X-\mu| \geq k\sigma] \leq \frac{1}{k^2} \iff P[|X-\mu| < k\sigma] \geq 1 - \frac{1}{k^2}$$

예를 들어, 확률변수 X의 평균이 $\mu = 25$이고 분산이 $\sigma^2 = 16$이라 하자.

이 경우 X가 17과 33 사이의 값을 가질 확률의 하한

$$P(17 \leq X \leq 33) = P(|X-25| \leq 8) = P(|X-\mu| \leq 2\sigma) \geq 0.75$$