확률 분포

FAST CAMPUS SCHOOL 2018
Copyright FAST CAMPUS Corp. All Rights Reserved

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

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 $x_1, \dots$ 에 대하여 $f(x_i) \geq 0$이며 $\sum_i f(x_i) = 1$이다.

이산형 확률변수 X의 확률밀도함수 $f(x)$는 $P(X = x) = f(x)$를 만족.

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

확률밀도함수

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.

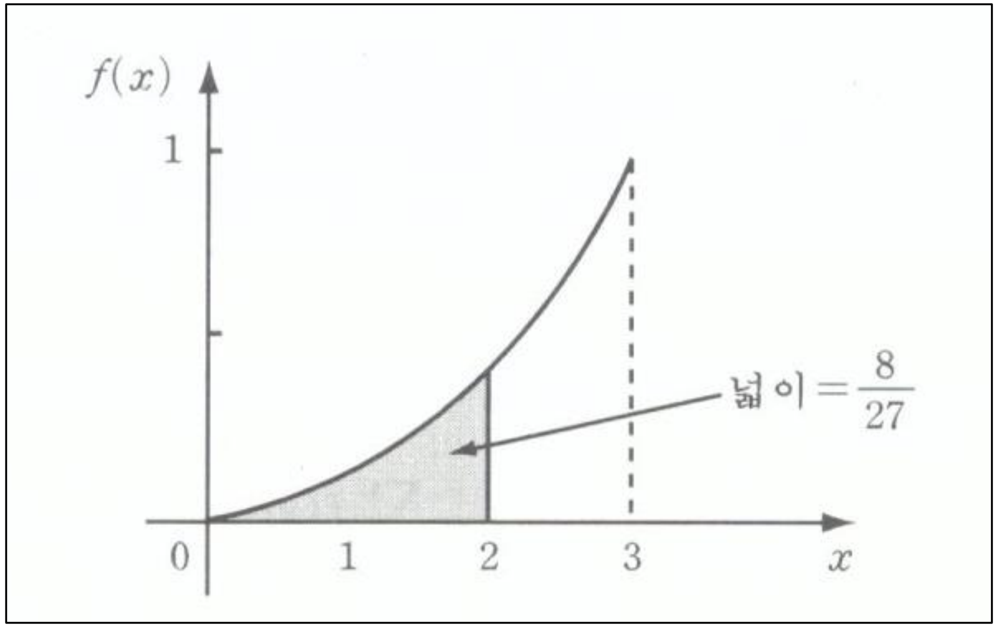
$\int_a^b f(x)dx = P(a \leq x \leq b)$와 같이 구간에 대한 확률값을 정의할 수 있다.

확률밀도함수 예제

- 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

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

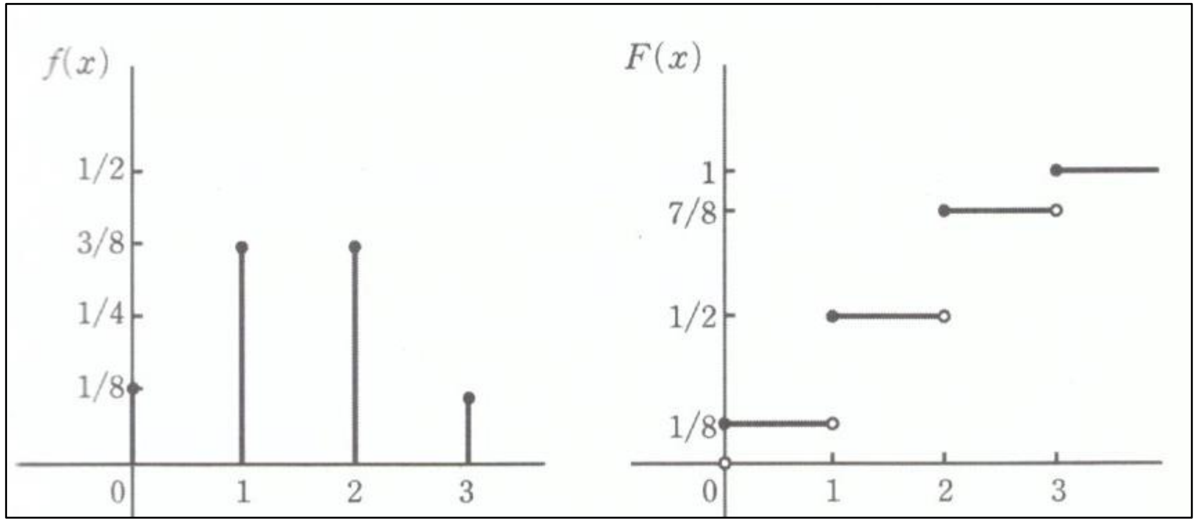
- 동전을 3회 독립반복하여 던졌을 경우 관심있는 변수 X(=앞면의 수)
  - ✓ 확률질량함수

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

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

- ✓ 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,

- 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 =하얀공의 수
  - ✓ Y =검은 공의 수
  - ✓ 두 변수 X, Y의 결합 확률밀도함수$(x = 0,1,2,3\,;\ \ y = 0,1,2\,;\ \ 0 \le x + y \le 3)$:

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

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

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

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.

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

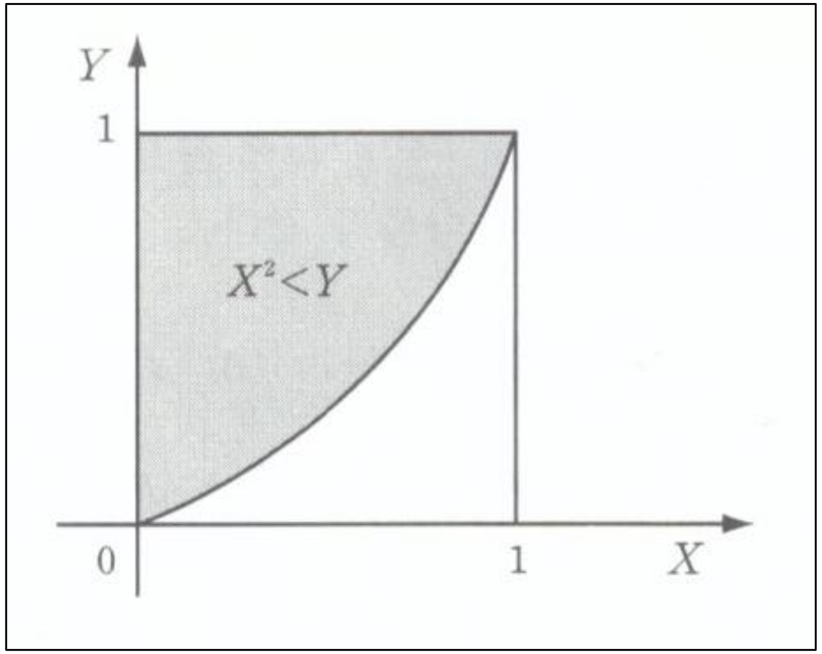
결합 누적분포함수 예제

- 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.

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

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

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

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 :
  - ✓ $\sum \dots \sum_{x_i}$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.

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

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

$\underline{f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)}$이므로 $\underline{X}$<u>와</u> $\underline{Y}$<u>는 독립</u>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

- 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 2

- 연속형 확률변수 X에 대하여
  - ✓ 확률밀도함수 :

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

  - ✓ $g(X) = X^2 + 5$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 변동 또는 흩어짐을 나타내는 측도로서, X가 평균 $E(X)$로부터 흩어져 있는(또는 밀집된) 정도
  - ✓ $\boldsymbol{Var(X) = E[X - E(X)]^2}$

- 표준편차 (standard deviation)
  - ✓ 분산의 제곱근
  - ✓ $\boldsymbol{\sigma_X = \sqrt{Var(X)}}$

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 $X_1, X_2, \ldots, X_n$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
- $Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big] = E(XY) - E(X)E(Y)$

공분산은 X와 Y의 선형관계에 대한 척도이다.

X가 증가할 때 Y가 증가하는 경향이 있으면 양의 값을,

X가 증가할 때 Y가 감소하는 경향이 있으면 음의 값을 가진다.

측정단위의 영향을 받는 단점이 있다.

체비셰프 부등식

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

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

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

이 경우 X가 17과 33 사이의 값을 가질 확률의 하한

$$\Downarrow$$

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

이산형 확률분포

- ✓ 베르누이 분포
- ✓ 이항 분포
- ✓ 포아송 분포
- ✓ 기하 분포
- ✓ 초기하 분포
- ✓ 음이항 분포

베르누이 분포(Bernoulli)

- 두 가지의 가능한 결과만을 갖는 시행(trial) 또는 실험(experiment)에서 정의되는 분포이다. 결과가 성공이면 1의 값을 가지고, 실패이면 0의 값을 가지는 확률변수를 베르누이 확률변수라고 한다.
  - ✓ 전화가 왔을 때, 전화를 한 사람이 여자인지 남자인지 측정
  - ✓ 주사위를 한 번 던졌을 때, 숫자 2가 나오는지를 측정
- 확률질량함수
  - $f(1) = p$
  - $f(0) = 1 - p = q$
- 기댓값과 분산
  - $E(X) = p$
  - $Var(X) = p(1 - p)$

이항 분포 (Binomial)

- 확률변수 $X_1, \dots, X_n$이 성공 확률이 p이며 서로 독립인 베르누이 시행으로부터 얻은 것일 때, 그들의 합으로 나타내어지는 $X = \sum_{i=1}^{n} X_i$의 분포이다. $X = \sum_{i=1}^{n} X_i$는 n회의 독립인 베르누이 시행에서 구한 '성공'의 횟수를 의미한다. 확률변수 X가 이항 분포를 따르는 것을 $X \sim B(n, p)$로 표현한다.
  - ✓ 한 축구 선수가 패널티킥을 차면 5번 중 4번을 성공한다고 한다. 이 선수가 10번의 패널티킥을 차서 7번 성공할 확률은?
  - ✓ A회사는 스마트폰의 한 부품을 만드는 회사로, 이 A사의 불량률은 5%로 알려져 있다. 이 회사의 제품 20개를 조사했을 때, 불량이 2개 이하로 나올 확률은?
- 확률질량함수
  - $f(x) = \binom{n}{x} p^x q^{n-x}$ (단, $x = 0, 1, \dots, n$)
- 기댓값과 분산
  - $E(X) = np$
  - $Var(X) = np(1-p)$

이항 분포 예제

- 어떤 주머니에 r개의 빨간 공과 w개의 하얀 공(단, $r + w = N$)이 들어 있다. 이제 n개의 공을 무작위 복원추출하였을 때

  ✓ 각 시행에서 빨간 공을 추출할 확률 :

  $$p = \frac{r}{N}$$

  ✓ X : 추출된 n개의 공 가운데 빨간 공의 개수

  ✓ X의 확률밀도함수 :

  $$f(x) = \binom{N}{n}\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)^{n-x}, \text{ (단, } x = 0,1,2,\ldots,n)$$

다항 분포 (multinomial)

○ k개의 상호배반인 사건 A_i에 대하여, $\bigcup_{i=1}^{k} A_i = S$이며, $P(A_i) = p_i$ $(i = 1, \dots, k$, $\sum_{i=1}^{k} p_i = 1)$이다. 실험을 n회 독립 반복시행했을 때 사건 A_i가 일어날 횟수를 X_i $(0 \le x_i \le n$, $\sum_{i=1}^{k} x_i = n)$라고 하면, 확률벡터 $\mathbb{X} = (X_1, X_2, \dots, X_k)$는 다항분포를 가진다고 한다.

○ $X_1, \dots, X_k$의 결합 확률밀도함수

- $f(x_1, \dots, x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$

다항 분포 예제

○ 5번의 동일한 게임으로 구성된 씨름 시합에서 두 선수 A,B 중 누구든지 먼저 3게임을 이기면 시합이 끝난다고 하자. 선수 A가 n번의 게임을 할 경우 이기는 횟수는 $B(n,\frac{1}{2})$를 따른다고 하자. 즉 매 게임에서 선수 A가 이길 확률은 $\frac{1}{2}$이고 게임 결과는 서로 독립이다. 이제 이 시합에서 선수 A가 이기는 횟수 X에 대한 기대값을 구해보자.

- 시합이 끝나기 위해 필요한 게임의 수 : N (=3,4,5)
- N의 확률밀도함수 :

$$f_N(n) = \binom{n-1}{2}\left(\frac{1}{2}\right)^{n-1}$$

- 기대값 $E(X)$:

$$E(X) = E_N\big(E(X|N)\big) = E_N\left(\frac{N}{2}\right) = \sum_{n=3}^{5}\frac{n}{2}f_N(n) = 2.0625$$

포아송 분포

시간이 지남에 따라 일어나는 어떤 특정한 사건의 발생횟수를 고려하자.

- 특정 시간동안 고속도로 구간에서 일어나는 대형사고의 횟수
- 특정 시간동안 전화교환대에 걸려오는 전화의 수
- 특정 시간동안 주유소에 주유를 위한 차량이 도착한 횟수

사건의 발생횟수에 대해 다음의 세 가지 성질이 성립한다고 가정하자.

- $\frac{\lim_{h\to 0} P(\text{길이가 } h \text{인 구간 내에서 사건 } A \text{가 } 1 \text{회 일어남})}{h} = \lambda$ (즉, 단위시간당 사건의 발생 확률)
- $\frac{\lim_{h\to 0} P(\text{길이가 } h \text{인 구간 내에서 사건 } A \text{가 } 2 \text{회 이상 일어남})}{h} = 0$
- 서로 겹치지 않는 두 구간 내에서 사건 A의 발생횟수는 서로 독립이다.

포아송 분포

○ 주어진 시간 $t > 0$에 대하여 확률변수 $X(t)$를 구간 $[0, t]$내에서 사건 A가 발생하는 횟수라고 하자. X가 아래와 같은 확률밀도함수를 가지고 사건의 발생횟수에 대한 세 가지 성질을 만족하면, 모수가 λ인 포아송(Poisson) 확률변수라 하고 $X \sim POI(\lambda)$로 표기한다.

○ X의 확률밀도함수

- $f(x) = \frac{\exp(-\lambda t)(\lambda t)^x}{x!} \quad \left(t = 1,\ f(x) = \frac{\exp(-\lambda)(\lambda)^x}{x!}\right)$

○ 기대값과 분산

- $E(X) = \lambda$
- $Var(x) = \lambda$

포아송 분포 예제

○ 어떤 전화교환대에서 매분 평균적으로 2건의 통화가 이루어진다고 한다. 통화횟수가 포아송 확률 과정을 따른다는 가정하에, 3분 동안에 5건 이상의 통화가 이루어질 확률은?

- 3분 동안에 걸려온 통화횟수 : 포아송 확률변수 X
- $\lambda = E(X) = 6$이다.
- $P(X \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{x=0}^{4} \frac{6^x e^{-6}}{x!} = 1 - 0.285 = 0.715$

기하 분포

- ○ "성공"확률이 p인 베르누이 시행을 독립적으로 반복할 때, 첫 번째 "성공"이 일어날 때까지 걸리는 시행횟수(X)를 나타내는 분포이다.
- ○ X의 확률질량함수
  - ● $f(x) = p(1-p)^{x-1} \ (x = 1,2,\ldots)$
- ○ X의 기대값과 분산
  - ● $E(X) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = \frac{1}{p}$
  - ● $Var(X) = \frac{1-p}{p^2}$
- ○ 기하확률변수는 무기억성을 갖는다.
  - ● $X \sim GEO(p)$이면, 임의의 자연수 j, k에 대하여 $P(X > j + k | X > j) = P(X > k)$

기하 분포 예제

○ 매주 발행되는 어떤 복권의 당첨확률이 0.001이고, 복권이 당첨되는 사건들은 독립이라 하자. 어떤 복권구입자가 매주 이 복권을 살 때, 처음 당첨될 때까지 소요되는 구매횟수를 확률변수 X라고 하자.

- X의 확률밀도함수
  - ▷ $f(x) = 0.001(1 - 0.001)^{x-1}$, $(x = 1,2, \dots)$
- X의 기대값
  - ▷ $E(X) = \frac{1}{0.001} = 1000$
- X의 분산
  - ▷ $Var(X) = \frac{1-0.001}{(0.001)^2} = 999000$

초기하 분포

○ 어떤 주머니에 r개의 빨간 공과 w개의 하얀 공(단, $r + w = N$)이 들어 있고, 그 중에서 n개의 공을 무작위 비복원추출(sampling without replacement)하였을 때, 선택된 빨간 공의 개수를 확률변수 X 라고 하자. 확률변수 X의 확률밀도함수가 아래와 같을 때, X의 분포는 초기하 분포이다.

○ X의 확률밀도함수

- $f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \ x = \max(0, n - N + r), \dots, \min(r, n)$

○ X의 기댓값과 분산

- $E(X) = \frac{nr}{N}$
- $Var(X) = \frac{nr}{N}\left[\frac{(N-r)(N-n)}{N(N-1)}\right]$

초기하 분포 예제

- 어떤 상자에 100개의 전자제품이 포장되어 있다. 한 구입자가 이 상자에 10개의 부품을 비복원 랜덤추출하여 불량품의 상태를 조사하려고 한다. 실제로 100개들이 상자속에 30개의 불량품이 있을 때, 2개 이하의 불량품이 나올 확률을 구해 보자.
  - 확률변수 X : 불량품의 개수
  - $X \sim HYP(10,100,30)$
  - $P(X \leq 2) = \sum_{x \leq 2} \frac{\binom{30}{x}\binom{70}{10-x}}{\binom{100}{10}} = 0.372857$

음이항 분포

- "성공"확률이 p인 베르누이 시행을 독립적으로 반복할 때, r개의 "성공"을 얻을 때까지 필요한 시행 횟수를 확률변수 X라고 하자. X의 확률밀도함수가 아래와 같을 때, 변수 X의 분포는 음이항 분포이다.(negative binomial)
- X의 확률질량함수
  - $f(x;r,p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$, $(x = r, r+1, \dots)$
- X의 기댓값과 분산
  - $E(X) = \frac{r}{p}$
  - $Var(X) = \frac{rq}{p^2}$

음이항 분포 예제

○ 두 개의 팀 A와 B가 겨루는 7회의 동일한 게임으로 구성된 경기에서 4회를 먼저 이기는 팀이 우승을 하게 된다. 어떤 팀이든지 먼저 4회를 이기면 경기는 더 이상 계속되지 않고 종료된다. A팀이 각 게임에서 이길 확률을 0.7이라고 할 때, 경기가 5회에서 종료될 확률을 구해 보자.

✓ A팀이 5회째 경기에서 4번째 승리를 거둘 사건 : A_5

✓ B팀이 5회째 경기에서 4번째 승리를 거둘 사건 : B_5

- $P(A_5) = \binom{4}{3}(0.7)^4(0.3) = 0.28812$
- $P(B_5) = \binom{4}{3}(0.3)^4(0.7) = 0.02268$
- 따라서 경기가 5회에서 종료될 확률은 0.28812+0.02268=0.3108이다.

연속형 확률분포

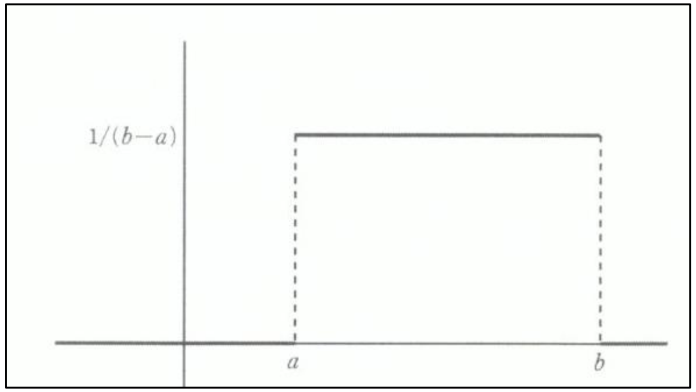
- ✓ 균일 분포
- ✓ 지수 분포
- ✓ 정규 분포
- ✓ 감마 분포
- ✓ 베타 분포

균일 분포

○ 확률변수 X가 실구간 (a,b)상에 균일(uniform)하게 분포되어 있을 때, 아래의 확률밀도함수를 가진다. X가 균일 분포를 따르면 $X \sim U(a,b)$로 표기한다.

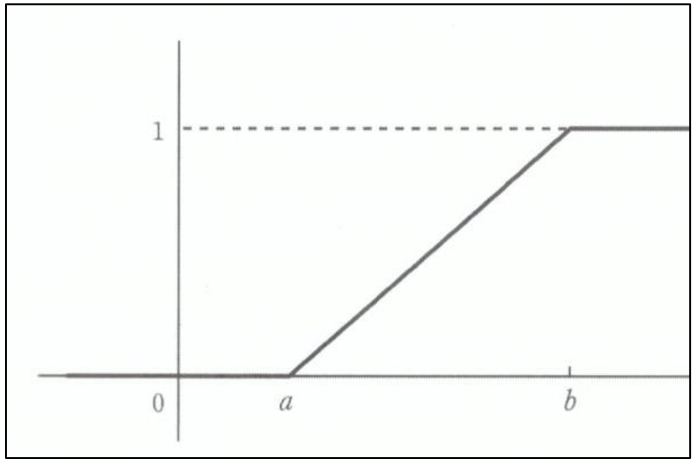
○ X의 확률밀도함수

- $f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{그 외의 경우} \end{cases}$



○ X의 분포누적함수

- $F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1, & x \ge b \end{cases}$



○ X의 기댓값와 분산

- $E(X) = \frac{b+a}{2}$
- $Var(X) = \frac{(b-a)^2}{12}$

지수 분포 예제

○ 지수분포는 포아송 확률변수를 고려했을 때, 특정한 사건 A가 일어나고 다음에 또 다시 같은 사건이 일어날 때까지 걸리는 시간 X(음이 아닌)을 나타내는 분포이다. 단위구간 내에서 평균발생횟수가 θ인 포아송 과정의 사건 사이 소요시간의 평균값은 $\lambda = 1/\theta$가 된다.

○ X의 확률밀도함수

- $f(x) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda})$, $(x > 0)$

○ X의 누적분포함수

- $F(x) = 1 - \exp(-\frac{x}{\lambda})$

○ X의 기댓값과 분산

- $E(X) = \lambda$
- $Var(X) = \lambda^2$

지수 분포 예제

○ 어떤 화학물질에서 α-입자가 10초당 평균 5개씩 포아송 과정을 따라 발생한다고 하자. 이때 첫 번째 입자가 발생될 때까지 걸리는 시간 X가 5초 이상일 확률을 구해보자.

✓ 초당 발생될 α-입자의 수의 기댓값은 $\theta = \frac{1}{2}$(즉, $\lambda = 2$)

✓ X의 확률밀도함수 :

$$f_X(x) = \left(\frac{1}{2}\right) e^{-\frac{x}{2}}, \ (x \geq 0)$$

✓ X가 5초 이상일 확률 :

$$P(X \geq 5) = \int_5^\infty \left(\frac{1}{2}\right) e^{-\frac{x}{2}} dx = e^{-\frac{5}{2}} = 0.08208$$

지수 분포 무기억성

○ 지수분포의 무기억성

- $X \sim EXP(\lambda)$이면, 양의 실수 a와 t에 대해서, $P(X > a + t | X > a) = P(X > t)$
- 예를 들어, 확률변수 X가 어떤 기계부품의 수명이라고 하면, $P(X > a + t | X > a)$는 시점 a에서 기계부품의 고장이 없을 때, 최소한 시간 t만큼 더 고장이 없을 사건에 대한 확률이다. 무기억 성질은 변수 X가 시점 a에서 기계부품의 고장이 없다는 조건을 "기억"하지 않고, 앞으로 시간 t만큼 더 고장이 없을 것만 고려한다는 것을 뜻한다. 즉, a시간만큼 일한 기계부품이 앞으로 t시간만큼 더 작동하는 확률이나 새 기계부품이 앞으로 t시간만큼 더 작동하는 확률이나 같다는 의미이다.

정규 분포

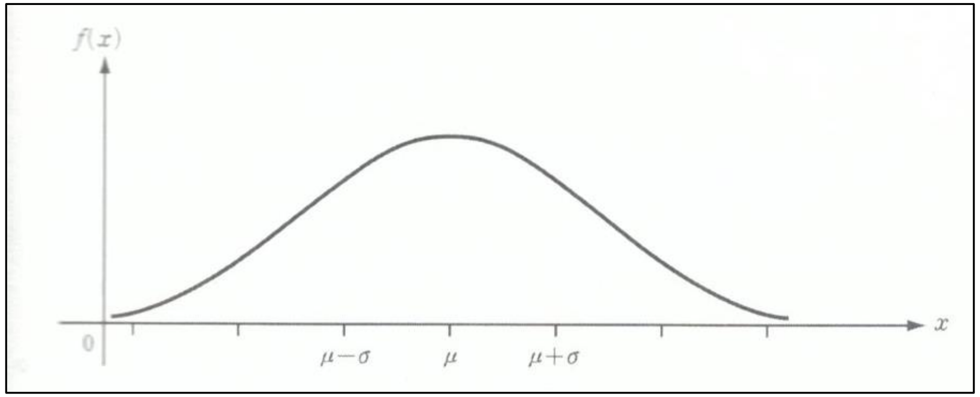
- 통계적 방법에서 가장 많이 이용되는 확률분포이며, 확률변수 X의 확률밀도함수가 다음과 같이 주어지면 X가 정규 분포(normal distribution)를 따른다고 한다.

- 확률밀도함수 :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

- X의 기대값과 분산
  - $E(X) = \mu$
  - $Var(X) = \sigma^2$

- X가 정규분포를 따르면 $X \sim N(\mu, \sigma^2)$로 표기하고, $Y = aX + b$에 대하여, $Y \sim N(a\mu + b, a^2\sigma^2)$로 나타낼 수 있다.



표준 정규 분포

○ 정규분포를 따르는 확률변수 X에 대하여 , $\boldsymbol{Z = \frac{X-\mu}{\sigma}}$라는 표준화 변환을 하면, Z의 확률밀도함수는 다음과 같이 주어진다. 이것을 표준 정규 분포(standard normal distribution) 라고 한다.

○ 확률밀도함수 :

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$$

○ Z가 표준 정규 분포를 따르면 $Z \sim N(0,1)$이라고 한다.

○ Z의 누적분포함수 :

$\Phi(x) = P(Z \leq x) = \int_{-\infty}^{x} \phi(t)dt$ 라고 할 때, X의 누적분포함수는 $F(x) = \Phi(\frac{x-\mu}{\sigma})$

이변량 정규 분포

- 두 변수 X,Y 의 결합 정규 분포 (bivariate normal distribution)이다.
- X의 결합 확률 밀도 함수 :

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

- X가 이변량 정규 분포를 따를 경우, $X \sim BVN(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$라고 표현한다.
- X, Y의 기댓값, 분산
  - $E(X) = \mu_X$, $E(Y) = \mu_Y$, $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$
- 공분산
  - $Cov(X,Y) = \rho\sigma_X\sigma_Y$

감마 분포

- 지수분포 등의 여러 가지 분포를 포함하는 분포족(family of distributions)의 하나이다. 감마 분포를 알기 위해서는 감마 함수를 알아야 한다.
- 감마 함수
  - $k > 0$에 대하여, $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt$
  - 감마함수의 특징
    - ✓ $\Gamma(1) = \int_0^\infty e^{-t}dt = 1$
    - ✓ $\Gamma(k) = (k-1)\Gamma(k-1)$, $k > 1$에 대하여
    - ✓ $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

감마 분포

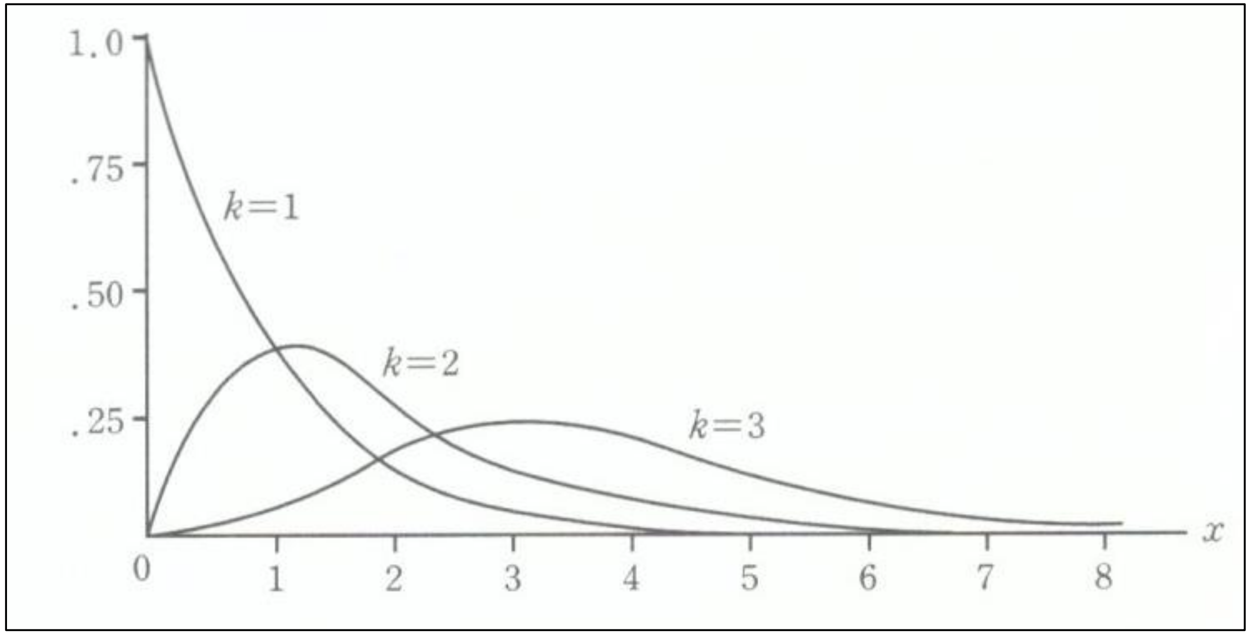
○ $\theta > 0$, $k > 0$, $x > 0$ 에 대하여, 확률변수 X가 아래의 확률밀도함수를 가지면, 감마분포라고 한다.

○ X의 확률밀도함수 :

$$f(x;\theta,k) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} \exp(-\frac{x}{\theta})$$

○ X의 기댓값과 분산

- $E(X) = k\theta$
- $Var(X) = k\theta^2$



베타 분포

○ 베타 함수 :

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dt$$

○ 확률변수 X가 다음의 확률밀도함수를 가지면, 베타분포라고 한다. 단 x의 값은 $0 < x < 1$ 이다.

$$f(x;\theta,k) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

○ X의 기댓값과 분산

- $E(X) = \frac{a}{a+b}$
- $Var(X) = \frac{(a+1)a}{(a+b+1)(a+b)} - \left(\frac{a}{a+b}\right)^2 \left(\frac{ab}{(a+b+1)(a+b)^2}\right)$