제 3장. 확률과 확률분포

확률의 정의

- **확률 (probability)** : 어떤 사건이 일어날 가능성의 수리적인 척도

- **실험 (experiment)** : 필요한 결과(outcome)를 얻기 위해 행하는 행위
- **표본공간 (sample space)** : 어떤 시행에서 모든 가능한 결과들의 집합
  - 모든 원소를 포함(exhaustive)
  - 상호배반(mutually exclusive)
- **단순사건(simple event)** : 발생 가능한 결과 (한가지)
- **사상(event, 사건)** : 관심있는 실험 결과들의 집합으로서 표본공간의 부분집합

(예 1) - 주사위를 던지는 실험에 대한 표본 공간
- 주사위의 눈이 짝수인 사건

(예 2) - 동전 두 개를 던지는 실험에 대한 표본 공간
- 첫번째 동전이 앞면이 나오는 사건

확률의 정의

- 사건의 연산 : 사건 A, B 에 대하여
  - 합사건 : $A \cup B$
  - 곱사건 : $A \cap B$
  - 여사건 : A^c
  - $A \cap B = \varnothing \Rightarrow$ A,B는 서로 배반(mutually disjoint)

- **확률의 고전적 정의** (Laplace; 1749-1827)

  : 각각의 근원사건이 일어날 가능성을 동일하게 정의

N 개의 원소로 구성된 표본공간 $S = \{e_1, e_2, \ldots, e_N\}$ 에서 각각의 근원사건이 일어날 가능성이 같은 경우에 m 개의 원소로 구성된 사건 A 의 확률은 $$P(A) = \frac{m}{N}$$ 으로 정의한다

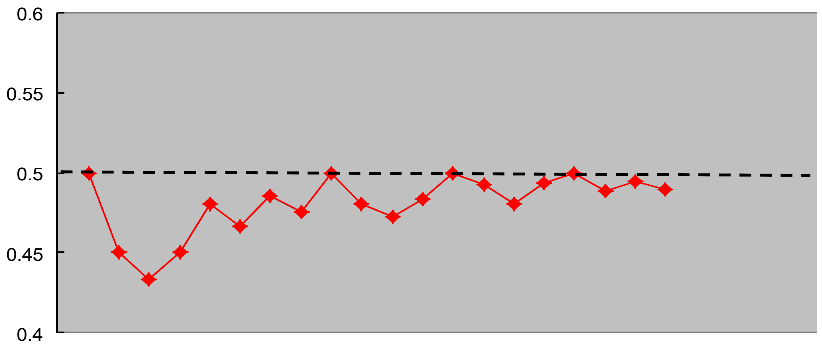
: 표본공간이 유한개의 원소만을 가질 때 적용될 수 있음

: 각 결과가 나타날 가능성이 같지 않은 경우도 있음

: 따라서 좀 더 일반적인 확률의 정의가 필요하게 됨

확률의 공리적 정의

- (예) 동전의 앞면이 나올 확률 = ½



- **확률의 공리적 정의** (Kolmogorov; 1903-1987)

  : 확률을 상대도수의 극한 개념으로 파악

(1) 표본공간 S 에서 임의의 사건 A 에 대하여 $0 \le P(A) \le 1$

(2) $P(S) = 1$

(3) 서로 배반인 사건 $A_1, A_2, \ldots$ 에 대하여

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

를 만족할 때, $P(A)$ 를 사건 A 의 확률이라 한다

조건부 확률과 곱셈법칙

- **조건부 확률 (conditional probability)**

  : 사건 A가 일어났다는 조건 하에 사건 B가 일어날 확률

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \qquad P(A) > 0$$

- **곱셈법칙**

  $P(A) > 0, \quad P(B) > 0$ 이면,

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

  - 예제 3.3 : 두 개의 주사위를 던지는 실험에서 첫 번째 던진 주사위의 눈이 두 번째 주사위의 눈보다 클 때, 두 주사위의 눈의 합이 10일 확률을 구하여라

  - 불량품 20개와 양호품 80개로 구성된 로트에서 2개의 제품을 단순랜덤추출할 때, 2개 모두 불량품일 확률을 구하여라

베이즈 정리

- **전확률공식**

  : 사건 $A_1, A_2, ..., A_n$ 에 대하여 $A_i \cap A_j = \varnothing \ (i \neq j), \quad A_1 \cup A_2 \cup ... \cup A_n = S, \ P(A_i) > 0$ 이면

$$P(B) = P(B \mid A_1)P(A_1) + ... + P(B \mid A_n)P(A_n)$$

- **베이즈 정리 (Bayes' Theorem)**

  : 표본 공간 S의 분할 $A_1, A_2, ..., A_n \ (P(A_i) > 0, \ i = 1, ..., n)$ 과 사건 $B(P(B) > 0)$ 에 대하여

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{k=1}^{n} P(B \mid A_k)P(A_k)}$$

  - 예) 어떤 지역의 결핵환자의 비율이 0.001로 알려져 있다. 결핵에 걸려있는지를 알아보는 검사에서 결핵에 걸렸을 때, 양성 반응이 나타날 확률은 0.95이고 그렇지 않을 때 양성반응이 나타날 확률이 0.011이라고 한다. 양성반응이 나타났을 때 결핵에 걸렸을 확률을 구하여라

독립사건

- 사건 A와 사건 B는 **서로 독립(mutually independent)**

$$\Leftrightarrow \quad P(B \mid A) = P(B)$$

- 사건 A와 B가 독립이면 $P(A \cap B) = P(A)P(B)$

(참고) A와 B가 서로 독립이면 A^C 와 B도 서로 독립이다

- 사건 $A_1, A_2, \ldots, A_n$ 에 대하여 다음이 성립 할 때, 사건 $A_1, A_2, \ldots, A_n$ 는 서로 독립이다

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad (1 \le i < j \le n)$$

$$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k) \quad (1 \le i < j < k \le n)$$

$$\ldots$$

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = \prod_{i=1}^{n} P(A_i)$$

확률변수

- **확률 변수 (random variable)**

  : 표본공간에서 정의된 실수 함수

  : 관찰이나 실험의 결과물에 대한 숫자적 표현 (자료처리의 편리성)

  - 예) 동전을 1개 던져서 나오는 결과를 관찰하는 실험

- **이산형 확률 변수** : 유한개이거나 셀 수 있는 값을 갖는 확률변수

  \- X=주사위의 눈의 개수

- **연속형 확률 변수** : 무한개의 셀 수 없는 가능한 값을 갖는 확률변수

  \- Y=전구의 수명

이산형 확률 분포

- **확률분포(probability distribution)**
  : 확률변수가 취할 수 있는 값들에 확률이 대응되어 있는 것

- **이산형 확률분포 (discrete probability distribution)**
  : 이산형 확률변수의 확률분포
  : 예) X = 주사위를 던졌을 때의 결과

- 확률분포표

이산형 확률 분포

- **확률분포함수 (probability distribution function)**
  : 확률변수의 확률분포를 함수로 나타낸 것

- 이산형 확률 변수의 **확률질량함수** (probability mass function)
  : 이산점에서의 확률을 나타냄

$$p(x) = \begin{cases} P(X = x_i), & x = x_i \ (i = 1, 2, \ldots, n) \\ 0 & , \ o.w. \end{cases}$$

- 확률질량함수의 성질
  - $0 \le p(x) \le 1, \qquad \sum_{all\ x} p(x) = 1$
  - $P(a < X \le b) = \sum_{a < x \le b} p(x)$

- **누적확률함수** (cumulative distribution function : cdf)
  : 확률변수 X가 x 보다 작거나 같을 확률

$$F(x) = P(X \le x) = \sum_{X \le x} p(x)$$

이산형 확률 분포

- 예제 3.7 : 어느 백화점에서 연말 매출용으로 내놓은 15개의 시계겸용 라디오 가운데 5개의 불량품이 있다고 하자. 한 고객이 이 가운데 3개를 단순랜덤추출 하였을 때, 그 중 불량품의 개수를 X 라고 하자. 확률 변수 X 의 확률분포를 구하여라

연속형 확률 분포

- 연속형 확률변수

: 확률변수가 취할 수 있는 값의 수가 셀 수 없이 많을 때

- 연속형 확률 변수의 **확률밀도함수**(probability density function)

: 면적이 확률을 나타냄

$$P(a \le X \le b) = \int_a^b f(x)dx$$

- 확률 밀도함수의 성질
  - $f(x) \ge 0, \quad \int_{-\infty}^{\infty} f(x)dx = 1$
  - 임의의 상수에 대하여 $P(X = a) = \int_a^a f(x)dx = 0$

$$\Rightarrow P(a < X < b) = P(a < X \le b) = P(a \le X < b) = P(a \le X \le b) = \int_a^b f(x)dx$$

- 연속형 확률변수의 누적분포함수

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\, dx$$

연속형 확률 분포

- 예제 3.4 : 확률변수 X는 [0,1]의 모든 값을 취할 수 있음

- 예제 3.8

$$p(x) = \begin{cases} bx(1-x), & 0 \le x \le 1 \\ 0 \quad , & o.w. \end{cases}$$

확률변수의 기대값

- 확률변수의 평균 (기대값)

$$E(X)=\begin{cases}\sum_{all\ x} xp(x) & \text{(이산형)}\\ \int_{-\infty}^{\infty} xp(x)dx & \text{(연속형)}\end{cases}$$

- 확률변수의 함수의 기대값

$$E\left[g(X)\right]=\begin{cases}\sum_{all\ x} g(x)p(x) & \text{(이산형)}\\ \int_{-\infty}^{\infty} g(x)p(x)dx & \text{(연속형)}\end{cases}$$

- 기대값의 성질
  - $E(aX+b)=aE(X)+b$
  - $E\left[c_1g_1(X)+c_2g_2(X)\right]=c_1E\left[g_1(X)\right]+c_2E\left[g_2(X)\right]$

확률변수의 분산과 표준편차

- 확률변수의 분산과 표준편차

$$Var(X) = E\left[(X-\mu)^2\right] = \begin{cases} \sum_{all\ x}(x-\mu)^2 p(x) & \text{(이산형)} \\ \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx & \text{(연속형)} \end{cases}$$

$$sd(X) = \sqrt{Var(X)}$$

- 분산의 간편계산법

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

- 분산의 성질
  - $Var(X) \geq 0$
  - $Var(aX+b) = a^2 Var(X)$

- 표준화 (standardization)

$$Z = \frac{X - E(X)}{sd(X)}$$

확률변수의 기대값과 분산 : 예제

- 예제 3.10 : X =동전을 2개 던졌을 때 나오는 앞면의 개수

x	
$p(x)$	

결합분포

- **결합 확률 분포 (joint probability distribution)**

  : 두 개의 확률변수가 취할 수 있는 값들의 모든 순서짝에 확률을 대응시켜 놓은 것

- 이산형 확률변수의 **결합확률밀도함수 (joint probability density function)**

$$p(x,y)=P(X=x,Y=y)$$

  - $0 \le p(x,y) \le 1$
  - $\sum_{all\ x}\sum_{all\ y} p(x,y)=1$
  - $P(a<X\le b,\ c<Y\le d)=\sum_{a<x\le b}\sum_{c<y\le d} p(x,y)$

- 연속형 확률변수의 **결합확률밀도함수**

$$P(a<X<b,\ c<Y<d)=\int_a^b\int_c^d f(x,y)dydx$$

  - $0 \le f(x,y)$
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy=1$

주변분포

- **주변확률밀도함수 (marginal probability density function)**

: 두 확률변수 X, Y의 결합분포에서 X만의 분포 또는 Y만의 분포

$$p_1(x) = \sum_{all\ y} p(x, y), \qquad p_2(y) = \sum_{all\ x} p(x, y)$$ (이산형)

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)dy, \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y)dx$$ (연속형)

[예 1]

$y \backslash x$	1	2	3
1	1/12	1/6	0
2	0	1/9	1/5
3	1/18	1/4	2/15

[예 2] $f(x, y, z) = \dfrac{xyz}{108}, \qquad x = 1, 2, 3;\ y = 1, 2, 3;\ z = 1, 2$

두 확률변수의 함수의 기대값과 독립성

- 두 확률변수 X, Y 의 결합확률밀도함수를 $p(x, y)$ 라고 할 때, 이들의 함수 g, g_1, g_2 에 대하여 다음의 사실들이 성립한다

(1) $E[g(X,Y)] = \sum_{all\ x}\sum_{all\ y} g(x,y)p(x,y)$

(2) $E[c_1 g_1(X,Y) + c_2 g_2(X,Y)] = c_1 E[g_1(X,Y)] + c_2 E[g_2(X,Y)]$

- 두 확률변수 X 와 Y 가 **서로 독립(mutually independent)** 이다

$$\Leftrightarrow \begin{cases} p(x,y) = p_1(x)p_2(y) & \text{(이산형)} \\ f(x,y) = f_1(x)f_2(y) & \text{(연속형)} \end{cases}$$

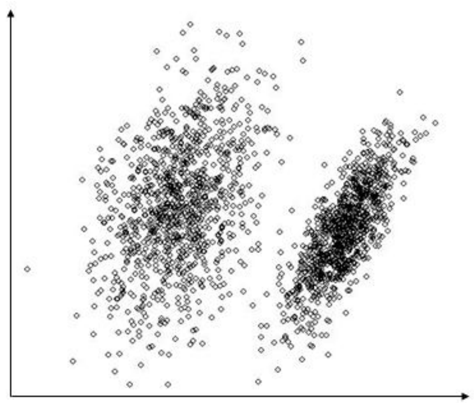
- (예) $f(x,y) = \begin{cases} \dfrac{3x-y}{9}, & 1 < x < 3,\ 1 < y < 2 \\ 0, & o.w. \end{cases}$

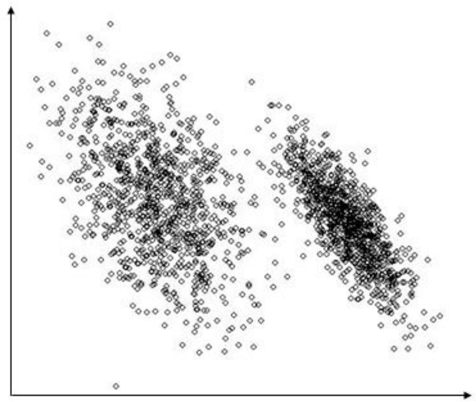
공분산

- **공분산 (covariance)**
  : 두 변수가 어느 방향으로 얼마나 변동하는지 또는 퍼져 있는지를 나타냄
  : 부호 (sign) 와 크기 (magnitude)
  : 자료의 단위에 영향을 받음

$$Cov(X,Y) = E\left[\left(X-\mu_x\right)\left(Y-\mu_y\right)\right] = E(XY) - E(X)E(Y)$$

- 성질
  - $Cov(X,Y) > 0$: 양의 선형 관계, x와 y가 평균에 대하여 서로 같은 값을 가질 때
  - $Cov(X,Y) < 0$: 음의 선형 관계, x와 y가 평균에 대하여 서로 반대 값을 가질 때
  - $Cov(X,Y) = 0$: 선형 관계 없음


양의 공분산의 크기 비교


음의 공분산의 크기 비교

상관계수

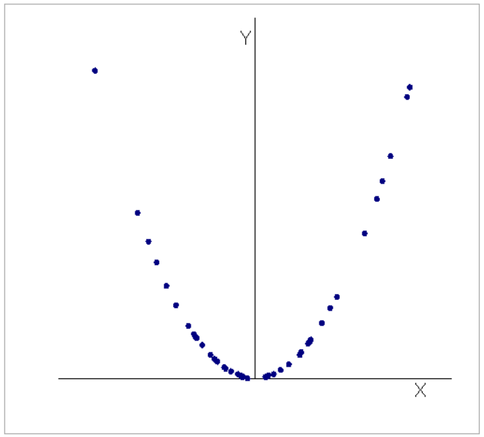
- **상관계수 (correlation coefficient)**

  : 공분산을 각자의 표준편차로 나누어 표준화 한 것

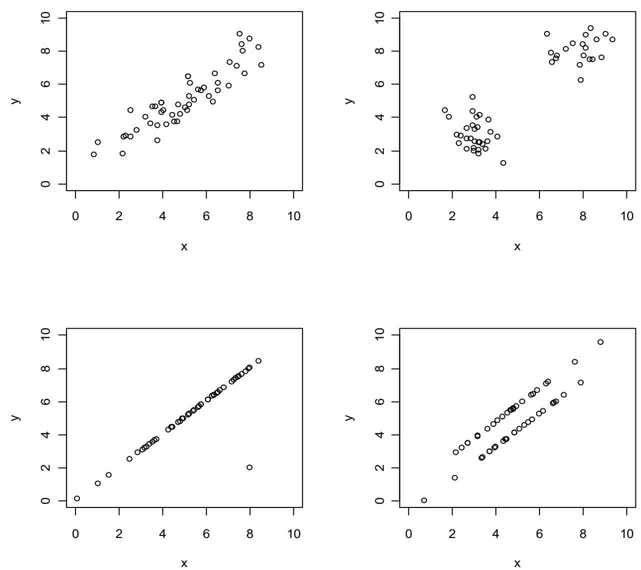
  : X,Y의 단위에 무관하게 연관성의 방향(부호)과 정도(크기)를 나타냄

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- 성질
  - $-1 \le \rho \le 1$
  - $\rho > 0$: 양의 선형관계
  - $\rho < 0$: 음의 선형관계
  - $\rho = 0$: 선형관계 없음



< 동일한 상관계수 값을 갖는 경우>



공분산과 상관계수의 성질

- $Cov(aX+b, cY+d) = acCov(X,Y)$

- $Corr(aX+b, cY+d) = \begin{cases} Corr(X,Y) & ,ac>0 \\ -Corr(X,Y) & ,ac<0 \end{cases}$

- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$

- $Var(X) = Cov(X,X)$

- $Y = aX + b \ (a \neq 0) \quad \Rightarrow \quad Corr(X,Y) = \dfrac{a}{|a|}$

확률변수의 독립성

- X, Y 가 독립일 때,
  - X 와 Y의 연관성은 없다
  - $E(XY) = E(X)E(Y)$ 이므로

$$Cov(X,Y) = 0, \qquad Corr(X,Y) = 0$$
$$Var(X+Y) = Var(X) + Var(Y)$$

  - $Cov(X,Y) = 0$ 가 X, Y가 독립임을 포함하지는 않는다

- X, Y가 독립이 아닐 때
  - $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$