Contents

Chapter 6 확률분포

01 확률변수

〈무조건 각각의 근원사건에 서로 다른 실수 값을 대응시킬 필요는 없다.〉

- 확률변수 (Random variable)
  : 각각의 근원사건들에 실수값을 대응시키는 함수, 표본공간에서 정의된 실수로의 함수
  X, Y, ... 등으로 표시

$$X : \Omega \rightarrow R$$

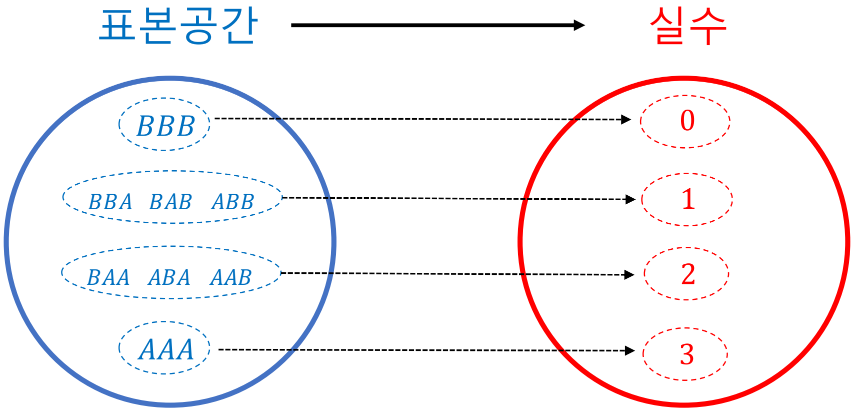
예) 동전을 2번 던지는 실험 : $\Omega = \{HH, HT, TH, TT\}$

예) 주사위를 던지는 실험 : $\Omega = \{1, 2, 3, 4, 5, 6\}$

① 이산확률변수 : 확률변수가 가질 수 있는 값들이 유한하거나 무한하더라도 셀 수 있는 경우
② 연속확률변수 : 연속적인 구간에 속하는 모든 값을 다 가질 수 있는 경우

Chapter 6 확률분포

예제1) 확률변수 X : 3명 중에서 A 회사 제품의 승용차를 소유한 사람의 수



$$P(X=0)=\frac{1}{8}, P(X=1)=\frac{3}{8}, P(X=2)=\frac{3}{8}, \quad P(X=3)=\frac{1}{8}$$

- 2.1, 2.2

02 이산확률변수와 확률분포

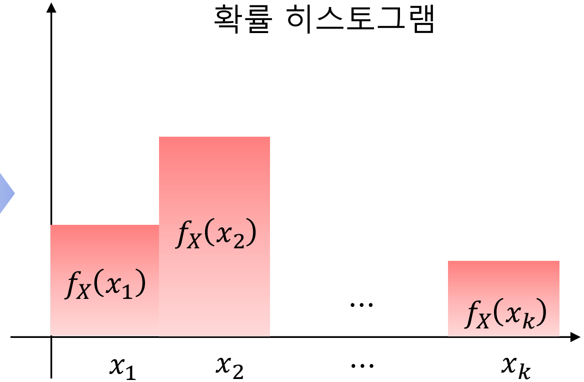
- 확률분포(Probability distribution)
 : 확률변수가 갖는 값들과 그에 대응하는 확률값을 나타낸 것으로 나열된 표나 수식으로 표현
 (보통은 확률변수 X의 분포라고 한다)

- 확률질량함수 (확률함수)
 : $f_X(x_i) = P(X = x)$
 모든 x_i 에 대해 $0 \leq f_X(x_i) \leq 1$ 이고 $\sum f_X(x_i) = 1$를 만족

확률분포(표)

X	$f_X(x) = P(X = x)$
x_1	$f_X(x_1)$
x_2	$f_X(x_2)$
$\vdots$	$\vdots$
x_k	$f_X(x_k)$
합계	1

확률 히스토그램



Chapter 6 확률분포

- 확률분포표의 예시

– 해당 함수 라라미터는 '확률 변수임'이라는 뜻.
즉, 각각의 확률 변수들이 인자값으로 들어온다는 뜻.

동전을 2번 던지기

X	$f_X(x) = P(X = x)$
0	1/4
1	1/2
2	1/4
합계	1

X : 동전을 던졌을 때, 앞면이 나오는 횟수

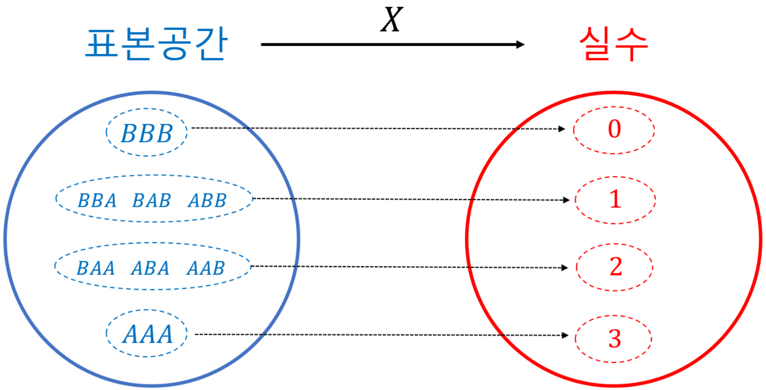
앞면이 나올 때까지 동전 던지기

X	$f_X(x) = P(X = x)$
1	1/2
2	1/4
3	1/8
4	1/16
⋮	⋮
합계	1

X : 동전을 던지는 횟수.

Chapter 6 확률분포

- 예제 3) 확률변수 X : 세 명 중 구두를 구매한 학생의 수



X	$f_X(x) = P(X = x)$	
0	$1/8$	← 0명이 구매한 확률
1	$3/8$	← 1명이 구매한 확률
2	$3/8$	← 2명이 구매한 확률
3	$1/8$	← 3명이 구매한 확률
~~4~~		
합계	1	

- 예제 4
- 예제 5
- 3.1, 3.20

03 확률분포의 기댓값과 표준편차

- 기댓값 (Random variable) : 확률분포의 모평균 (모집단의 평균)

$$E(X) = \mu_x \quad = \mu_x.$$

$$= \sum (\text{확률변수가 취하는 값}) \times (\text{그 값을 가질 확률})$$

$$= \sum x_i \times P(X = x_i) = \sum x_i \times f_X(x_i)$$

X	$f_X(x) = P(X = x)$
0	1/8
1	3/8
2	3/8
3	1/8
합계	1

$xf_X(x)$
$0 \cdot 1/8$
$1 \cdot 3/8$
$2 \cdot 3/8$
$3 \cdot 1/8$
$\mu_X = 1.5$

$$E(X) = \sum x_i \times f_X(x_i)$$

$$= 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1.5$$

$E(X)=1.5$, $\mu_X=1.5$)

Chapter 6 확률분포

결국, 분산은 표준편차를 구하기 위해 어쩔 수 없이 구해야하는 것이다.
(분산은 살짝 쩌리 느낌이 든다.)

- 분산 (Variance) : 확률분포의 모분산 (모집단의 분산)
  모평균으로부터의 편차 제곱의 기대값으로 정의됨

$$Var(X) = \sigma_X^2 == \sum (x_i - \mu_X)^2 \times f_X(x_i)$$

- 분산의 계산식

$$Var(X) = E(X^2) - \left(E(X)\right)^2 = \sum x_i^2 f_X(x_i) - \mu_X^2$$

- $E(X-\mu_X)^2 = E(X^2 - 2\mu_X X + \mu_X^2) = E(X^2) - 2\mu_X E(X) + \mu_X^2 = E(X^2) - \mu_X^2$

〈모집단 내 편차들의 기댓값〉

- 표준편차 (Standard Deviation) : 확률분포의 모표준편차 (모집단의 표준편차)
  모분산의 양의 제곱근

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

- 예제 8

- 예제 9

X	$f_X(x) = P(X = x)$	$xf_X(x)$	$X - \mu_X$	$(X - \mu_X)^2$	$(X - \mu_X)^2 f_X(x)$	$x^2 f_X(x)$
0	0.1	0	-2	4	0.4	0
1	0.2	0.2	-1	1	0.2	0.2
2	0.4	0.8	0	0	0	1.6
3	0.2	0.6	1	1	0.2	1.8
4	0.1	0.4	2	4	0.4	1.6
합계	1.0	2.0=μ_X	0	10	1.2=$\sigma^2{}_X$	5.2

$$\sum(x_i - \mu_X)^2 \times f_X(x_i) = 1.2\ ,\ E(X^2) - \mu_X^2 = 5.2 - 4 = 1.2$$

- 4.1, 4.5, 4.6, 4.11, 4.12

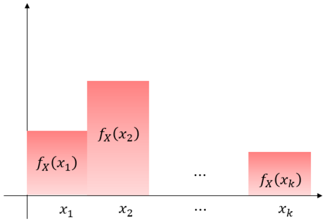
Chapter 6 확률분포

- 통계적추론과 확률분포

통계적 추론을 위해서는 먼저 모집단에 대해 분포가정을 해야 한다.
예) 어떤 자료(변수 X)가 "OO한 확률분포를 따른다" 라고 가정하는 것.

모집단의 확률분포
(참 확률분포)

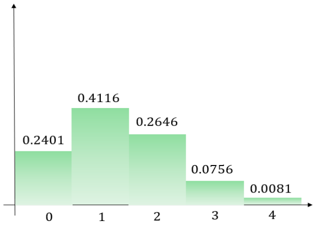
X	$f_X(x)$
x_1	$f_X(x_1)$
$\vdots$	$\vdots$
x_m	$f_X(x_m)$
Σ_x	1



모집단의 중심위치의 척도 : μ_X
모집단의 퍼진정도의 척도 : σ_X^2 (또는 , σ_X)

※ 이렇게 모집단의 특성을 나타내는 값을 모수 (parameter)라고 함

표본으로부터 얻어진
(모집단에 대한) 추정된 확률분포



표본의 중심위치의 척도 : $\bar{X}$
표본의 퍼진정도의 척도 : S_X^2 (또는 , s_X)

※ 모수에 대한 추론을 위해 표본으로부터 계산되는 양을 통계량(statistic)라고 함

04 두 확률변수의 결합분포

- 확률변수 X와 Y의 결합확률분포(Joint probability distribution)
  : X가 취하는 값과 Y가 취하는 값의 각 쌍에 대응되는 확률

$$\text{For } X=(x_1,\cdots,x_m),\quad Y=(y_1,\cdots,y_n),\quad i=(1,\cdots,m),\quad j=(1,\cdots,n)$$

$$f(x_i,y_j)=P(X=x_i,\quad Y=y_j)$$

	y_1	y_2	$\cdots$	y_n
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$	$\cdots$	$f(x_1,y_n)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$	$\cdots$	$f(x_2,y_n)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	$f(x_m,y_1)$	$f(x_m,y_2)$	$\cdots$	$f(x_m,y_n)$

- 예제 10

Chapter 6 확률분포

$X = \{x_1, x_2, x_3, x_4, \cdots, x_n\}$

- X와 Y의 주변확률분포(Marginal probability distribution)

: X와 Y의 결합확률분포로부터 다음의 식으로 계산된 개별 확률변수 X, Y의 확률분포

$$f_X(x_i) = P(X = x_i) = \sum_y f(x_i, y_j), \quad f_Y(y_j) = P(Y = y_j) = \sum_x f(x_i, y_j)$$

$y = 1, 2, 3, \cdots n$

	y_1	y_2	$\cdots$	y_n	Σ_x
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$\cdots$	$f(x_1, y_n)$	$f_X(x_1)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$\cdots$	$f(x_2, y_n)$	$f_X(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	$\cdots$	$f(x_m, y_n)$	$f_X(x_m)$
Σ_y	$f_Y(y_1)$	$f_Y(y_2)$	$\cdots$	$f_Y(y_n)$	1

X	$f_X(x)$
x_1	$f_X(x_1)$
$\vdots$	$\vdots$
x_m	$f_X(x_m)$
Σ_x	1

Y	$f_Y(y)$
y_1	$f_Y(y_1)$
$\vdots$	$\vdots$
y_n	$f_Y(y_n)$
Σ_y	1

- 예제 11

	$y=0$	1	2	3
$x=0$	0.05	0.05	0.10	0.00
1	0.05	0.10	0.25	0.10
2	0.00	0.15	0.10	0.05

$P(x)$

X	$f_X(x)$
0	0.2
1	0.3
2	0.5
Σ_x	1

$P(Y)$

Y	$f_Y(y)$
1	0.1
2	0.3
3	0.45
4	0.15
Σ_y	1

$P(X>Y)$

	$y=0$	1	2	3
$x=0$				
1	0.05			
2	0.00	0.15		

$P(X+Y=3)$

	$y=0$	1	2	3
$x=0$				0.00
1			0.25	
2		0.15		

- 확률변수 X와 Y의 결합확률분포로부터 X와 Y로 이루어지는 "새로운 확률변수 Z"에 대한 확률분포를 계산할 수 있음

$P(Z=X+Y)$

	0	1	2	3	4	5	Σ_z
$f_Z(z)$	0.05	0.10	0.20	0.40	0.20	0.05	1

$$E(Z)=\sum z_i \times f_Z(z_i)=2.75$$

Chapter 6 확률분포

- 두 확률변수의 합에 대한 기댓값

$$E(X) = \sum x_i \times f_X(x_i) \quad \text{and} \quad f_X(x) = \sum_y f(x_i, y_j)$$

$$E(a + bX + cY) = \sum_x \sum_y (a + bx_i + cy_j) f(x_i, y_j)$$

$$= a + b \sum_x \sum_y x_i f(x_i, y_j) + c \sum_x \sum_y y_j f(x_i, y_j)$$

$$= a + b \sum_x x_i \sum_y f(x_i, y_j) + c \sum_y y_j \sum_x f(x_i, y_j)$$

$$= a + b \sum_x x_i f_X(x_i) + c \sum_y y_j f_Y(y_j) = a + b \cdot E(X) + c \cdot E(Y)$$

시험에 잘 나온다.

※ $E(X+Y) = E(X) + E(Y)$

- 4.1, 4.5, 4.6, 4.11, 4.12

관계의 '세기'는 알 수 없다!

Chapter 6 확률분포

2직 분포에서 정보를 얻을 수 없다

결과 값에서 두 변수의 단위가 약분되지 않는다.

X가 커질때 Y가 커지는 경우 (ex) 키와 몸무게의 관계)

X가 커질 때 Y가 작아지는 경우 (ex) 게임 하는 시간과 성적의 관계)

양수 or 음수

05 공분산과 상관계수

- E(X) 는 상수임!!!
- $E(X) \Leftrightarrow \mu_X,\ \mu_X$
- E(특정상수) = 특정상수.

- 공분산 (Covariance) : 두 확률변수의 선형관계의 정도를 측정한 값

공동의 분산

$$Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y) = E(X-E(X))(Y-E(Y))$$

$$= E(XY-\mu_Y X-\mu_X Y+\mu_X\mu_Y) = E(XY)-\mu_Y E(X)-\mu_X E(Y)+\mu_X\mu_Y = E(XY)-\mu_X\mu_Y$$

$$= E(XY) - E(X)\cdot E(Y)$$

전개

$$E(XY) = \sum_x \sum_y x_i y_j f(x_i,y_j)$$

Σ 파이썬으로 치면, for 문을 두번 중첩한 것이다!

X, Y의 결합확률 분포표

- 예제 12

	$y=0$	1	2	3	Σ_x	
$x=0$	0.05	0.05	0.10	0.00	0.20	← P(X=0)
1	0.05	0.10	0.25	0.10	0.50	← P(X=1)
2	0.00	0.15	0.10	0.05	0.30	← P(X=2)
Σ_y	0.10	0.30	0.45	0.15	1.00	
	P(Y=0)	P(Y=1)	P(Y=2)	P(Y=3)		

$0\times0\times f(0,0) + 0\times1\times f(0,1) + 1\times0\times f(1,0) + \cdots + 2\times3\times f(2,3) = 1.9$

$$E(XY) = \sum_x \sum_y x_i y_j f(x_i,y_j) = 1.9$$

$$E(X) = \sum_x x_i f_X(x_i) = 1.10$$

$$E(Y) = \sum y_i f_Y(y_i) = 1.65$$

Chapter 6 확률분포

$cov(aX, bY) = ab \cdot cov(X,Y)$

$E(X^2) - \{E(X)\}^2$

$E(XY) - E(X) \cdot E(Y)$

$= E(X^2) - \{E(X)\}^2$

- $Var(X) = E(X-\mu_X)(X-\mu_X) = Cov(X,X)$
- $Var(aX+b) = E\{(aX+b) - E(aX+b)\}\{(aX+b) - E(aX+b)\}$
  $= E[a(X-\mu_X)][a(X-\mu_X)] = a^2 \cdot E(X-\mu_X)(X-\mu_X) = a^2 \cdot Var(X)$

〈덧셈 무시〉

- $Cov(aX+b, cY+d) = E\{(aX+b) - E(aX+b)\}\{(cY+d) - E(cY+d)\}$
  $= E[a(X-\mu_X)][c(Y-\mu_Y)] = ac \cdot E(X-\mu_X)(Y-\mu_Y) = ac \cdot Cov(X,Y)$

↳ 덧셈 무시 ⇒ 분산과 같은 성질!!

- $-\infty < Cov(X,Y) < +\infty$ and $Var(X) \geq 0$

???

- $Var(X \pm Y) = E((X \pm Y) - E(X \pm Y))^2 = E((X-\mu_X) \pm (Y-\mu_Y))^2$
  $= E(X-\mu_X)^2 + E(Y-\mu_Y)^2 \pm 2 \cdot E(X-\mu_X)(Y-\mu_Y)$
  $= Var(X) + Var(Y) \pm 2 \cdot Cov(X,Y)$

└ 무조건 '+' 연산이 쓰임

- Ex. $Cov(2X+3Y, X-Y) =$ $\begin{cases} Cov(2X, X) + cov(3Y, X) + cov(2X, -Y) \\ + Cov(3Y, -Y) \end{cases}$

〈4 세트의 ... 더해진 ...〉

$= 2 \cdot Var(X) + 3 \cdot Cov(X,Y) - 2 \cdot Cov(X,Y) - 3 \cdot Var(Y)$

cov가 ... 다항식으로 전개됨!

$= 2 \cdot Var(X) - 3var(Y) + cov(X,Y$

- (모) 상관계수 (Correlation coefficients) :
  두 확률변수의 선형관계를 상대적인 값으로 나타낸 것

corr(X, Y)에 대한 다른 기호!!

$\langle \Leftrightarrow$ '$n-1$'로 나눈 것과 같다. (표준화) $\rangle$

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho_{X,Y}, \quad -1 \le \rho_{X,Y} \le 1$$

└ 두 변수의 공분산을 두 변수의 표준편차 곱으로 나눈 것이다.

- $Corr(aX+b, cY+d) = \frac{Cov(aX+b, cY+d)}{\sqrt{Var(aX+b)} \cdot \sqrt{Var(cY+d)}}$

$$= \frac{ac \cdot Cov(X,Y)}{|a|\sigma_X \cdot |c|\sigma_Y} = \frac{ac}{|a||c|} \cdot \rho_{X,Y}$$

$\sigma(aX+b) = \sqrt{Var(aX+b)}$
$= \sqrt{a^2 \times Var(X)}$
$= |a| \times \sqrt{Var(X)} = |a| \times \sigma(X)$

< 상수의 영향을 받지 않음! > (곱해진 상수의 부호에만 영향을 받는다.)

└ "a, c"

→ '양의 관계'의 강도와 '음의 관계'의 강도까지 알 수 있다!

- 6.1, 6.4

. 이전 장의 상관계수는 표본집단 간의 상관계수이고
지금의 〃 모집단 간의 상관계수이다.

확률과 통계에서 '독립' 이라는 개념이 성립하면, 계산이 상대적으로 쉬워진다.

06 두 확률변수의 독립성

- 두 사건 A와 B가 서로 독립:

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

- 두 확률변수 X와 Y가 서로 독립: (X, Y가 서로 독립일 때,)

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

$$\leftrightarrow f(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j)$$

$$\text{for all } (x_i, y_j)$$

- 예제 13

Chapter 6 확률분포

두 확률변수 X, Y가 독립이라면,

If $X \perp Y$, then

① $E(XY) = \sum_x \sum_y x_i y_j f(x_i, y_j) = \sum_x \sum_y x_i y_j f_X(x_i) \cdot f_Y(y_j)$

$= \sum_x x_i f_X(x_i) \cdot \sum_y y_j f_Y(y_j) = E(X) \cdot E(Y)$

② $Cov(X,Y) = E(XY) - \mu_X \mu_Y = 0$ and $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0$

③ $Var(X \pm Y) = Var(X) + Var(Y) \pm 2 \cdot Cov(X,Y) = Var(X) + Var(Y)$

무조건 '+'로 바뀐다!!!

→ Cov(X,Y) → Corr(X,Y)

- 두 확률변수가 독립이면, 공분산과 상관계수는 0 이다.
- 공분산과 상관계수는 0 이라고 해서 두 확률변수가 반드시 독립은 아니다.

반대는 무조건 성립 X

필요충분 조건의 관계가 아니다!!!

- 예제 14
- 7.1