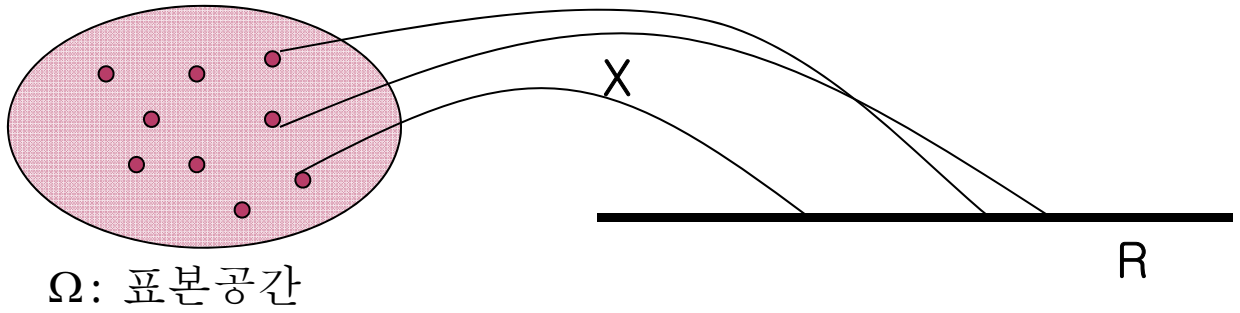
확률및통계

제6주 확률변수

hylee@silla.ac.kr

확률변수 (RANDOM VARIABLES)



- 각각의 근원 사건에 실수를 대응 시키는 함수
  - 동전 세 번 던지기에서 표면의 수, 어느 사거리에서 1시간 동안 지나가는 차의 대수
  - 사람의 키, 체온, 무게

◉ 이산확률변수 : 셀 수 있는 값들을 갖는 변수

- cf: 연속확률변수 : 연속적인 어느 구간 내의 값을 갖는 변수

◉ 확률분포 (Probability Distribution)

- 확률 변수가 갖는 값들과 이에 대응하는 확률을 나타낸 것 (표, 수식)

◉ 이산확률분포

- 이산확률변수 X가 취하는 각 값에 확률을 대응시킴으로써 구해진다.
- (eg. 예제3)

  X : 구두를 구매한 학생의 수 0,1,2,3 구두 : A 운동화 : B

X	경우의 사상	확률=P[X=x]=f(x)
0	BBB	1/8
1	ABB BAB BBA	3/8
2	AAB ABA BAA	3/8
3	AAA	1/8
		1

f : 확률함수,확률질량함수,
(확률밀도함수)

◉ 확률함수의 특성

- X: 이산확률변수 $x_1, x_2, \cdots, x_n$ 을 취한다.

  $f(x_i) = P[X = x_i]$ 는 다음을 만족시킨다.

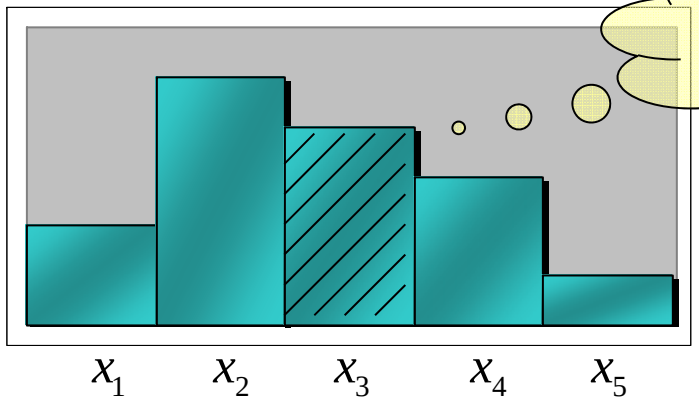
(i) 모든 x_i 에 대하여 $0 \le f(x_i) \le 1$

(ii) $\sum_{i=1}^{n} f(x_i) = 1$

◉ 확률 히스토그램

면적의 합 =1

- cf :



면적 = $f(x_3) = P[X = x_3]$

예제

- 예제 **4**

  -병충해 있음 **:30% - S**

  -병충해 없음 **:70% - F**

  4 그루를 임의 추출, **X :** 병충해가 있는 나무의 수

x	경우의 수	$f(x)$
0	FFFF(1)	1/16
1	FFFS,FFSF,FSFF,SFFF(4)	4/16
2	FFSS,FSFS,FSSF,SFFS,SFSF,SSFF(6)	6/16
3	FSSS,SFSS,SSFS,SSSF(4)	4/16
4	SSSS(1)	1/16
		1

경험적 확률분포

◉ 경험적 확률분포 (예제 5)

- X : 이메일 계정의수 $P(X=3)=?$ 미지수

  400명 조사

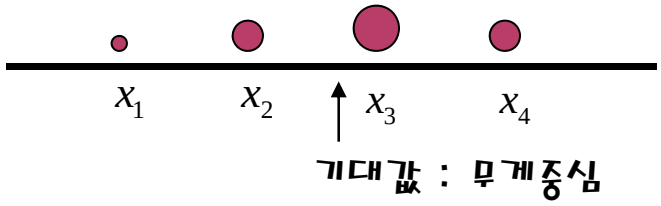
  $r_N(A) = A$의 상대도수

  $r_{400}(X=3) = 0.14 \approx P[X=3]$

➔ 경험적 확률분포 : 표본에 따라 변화한다.

기대값

- 분포의 무게 중심 ; 기대가치; ...



eg) 행운권의 기대상금

상금(만원)	장 수
x_1 : 1000	1
x_2 : 100	4
x_3 : 10	10
x_4 : 1	100
x_5 : 0	99,885
계	100,000

$$\text{기대상금} = \frac{\text{전체상금}}{\text{전체장수}}$$

$$= \frac{10,000,000\times 1+1,000,000\times 4+100,000\times 10+10,000+100}{100,000}$$

$$= 10,000,000\times\frac{1}{100,000}+1,000,000\times\frac{4}{100,000}$$
$$+100,000\times\frac{10}{100,000}+10,000\times\frac{100}{100,000}$$

$$= \sum_{i=1}^{5} x_i f(x_i) = 160\text{원}$$

- 구하는 식 :

X: 확률변수 $\quad x_1, x_2, \cdots, x_n$

$$E(X) = \mu_X = \mu \ : \ X\text{의 기대값}$$

$$= \sum_{i=1}^{n} x_i f(x_i)$$

eg

x	$f(x)$	$xf(x)$
0	0.1	0
1	0.2	0.2
2	0.4	0.8
3	0.2	0.6
4	0.1	0.4
		2

$E(X)$

eg) #4.3

X : 이익

입찰 O : +400,000,000 (확률 : 1/4)

입찰 X : -5,00,000 (확률 : 3/4)

$$\therefore E(X) = 400,000,000 \times \frac{1}{4} + (-5,000,000) \times \frac{3}{4} = 9625(\text{만원})$$

기대값의 특성

- h : 함수

$$E[h(X)] = \sum_{i=1}^{n} h(x_i) f(x_i)$$

- 일반적으로

$$E[h(X)] \neq h[E(X)]$$

- 예외: $h(X) = aX + b$ 선형함수

$$E(h(X)) = E(aX + b) = aE(X) + b = h(E(X))$$

표준편차 (STANDARD DEVIATION)

- 평균으로부터 퍼져있는 정도를 나타내는 수치.
- 분산 : cf : 표본분산 : $\frac{1}{n-1}\sum(x_i-\bar{x})^2$

$$Var(X)=E[(X-\mu)^2]=\sum(x-\mu)^2 f(x)=\sum x^2 f(x)-2\mu\sum xf(x)+\mu^2\sum f(x)$$
$$=E(X^2)-2\mu^2+\mu^2=E(X^2)-\mu^2$$

$$s.d(X)=\sqrt{Var(X)}=\sqrt{E[(X-\mu)^2]}$$

◉ 표준편차의 특성

- X : 확률변수 ; $E(X)=\mu$
- $a,\ b$: 상수

$$Var(aX+b)=a^2Var(X)$$
$$s.d.(aX+b)=|a|\,s.d.(X)$$

◉ 결합확률분포 (Joint Probability distribution)

- **eg. (**키, 몸무게**), (**중간성적, 기말성적**), (**성별, 교육정도**)**
- (X,Y) : 두 개의 확률변수에 관심
- (X,Y) 의 결합확률분포 : (X,Y)가 취하는 모든 값에 대응되는 확률을 제시함으로써 얻어진다.

* $X: x_1, x_2, \cdots, x_m$ 을 취한다.

$Y: y_1, y_2, \cdots, y_n$ 을 취한다.

X \ Y	y_1	y_2	$\cdots$	y_n	합 계
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$		$f(x_1,y_n)$	$\sum_{j=1}^{n} f(x_1,y_j) = f_X(x_1)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$		$f(x_2,y_n)$	$\sum_{j=1}^{n} f(x_2,y_j) = f_X(x_2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	$f(x_m,y_1)$	$f(x_m,y_2)$		$f(x_m,y_n)$	$\sum_{j=1}^{n} f(x_m,y_j) = f_X(x_m)$
합 계	$\sum_{i=1}^{m} f(x_i,y_1)$		$\cdots$	$\sum_{i=1}^{m} f(x_i,y_n)$	1
	$\Downarrow$			$\Downarrow$	
	$f_Y(y_1)$			$f_Y(y_n)$	

$$\sum_{i=1}^{m} f(x_i,y_1) = \sum_{i=1}^{m} P[X=x_i, Y=y_1]$$

$$= P[X=x_1, Y=y_1] + P[X=x_2, Y=y_1] + \cdots + P[X=x_m, Y=y_1]$$

$$= P\left[\bigcup_{i=1}^{n} (X=x_i, Y=y_1)\right] = P[Y=y_1] = f_Y(y_1)$$: 주변확률분포

기대값

- $g(X,Y)$: X,Y 의 함수

$$E[g(X,Y)] = \sum_{i=1}^{m}\sum_{j=1}^{n} g(x_i, y_j) f(x_i, y_j)$$

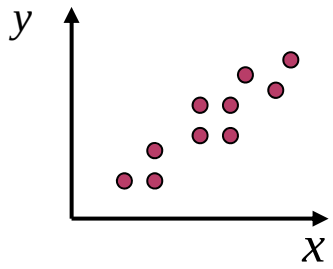
$$eg.\quad E(XY) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j f(x_i, y_j)$$

$$E(X) = \sum_i \sum_j x_i f(x_i, y_j) = \sum_i x_i \sum_j f(x_i, y_j) = \sum_i x_i f_X(x_i) = \mu_X$$

$$E(aX + bY) = \sum_i \sum_j (ax_i + by_j) f(x_i, y_j)$$

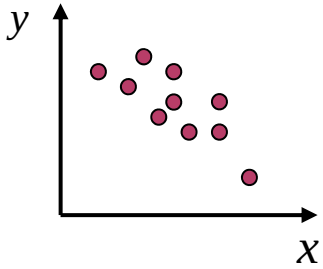
$$= a\sum_i \sum_j x_i f(x_i, y_j) + b\sum_i \sum_j y_j f(x_i, y_j) = aE(X) + bE(Y)$$

공분산과 상관계수 (COVARIANCE, CORRELATION COEFFICIENT)



x 증가 ➔ y 증가
x 감소 ➔ y 감소

양의 상관관계



x증가 ➔ y 감소
x 감소 ➔ y 증가

음의 상관관계

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

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

cf. $E(XY) = \sum_i \sum_j x_i y_j f(x_i, y_j)$

cf. $r = \dfrac{\sum x_i y_j - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_j^2 - n\bar{y}^2}}$

$= \dfrac{\frac{1}{n}\sum x_i y_j - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum y_j^2 - \bar{y}^2}}$

◉ 공분산과 상관계수의 특징

- a, b : 상수

$$Cov(aX, bY) = abCov(X, Y)$$

$$< E[(aX - a\mu_X)(bY - b\mu_Y)] = abE[(X - \mu_X)(Y - \mu_Y)] >$$

$$Corr(aX, bY) = \frac{Cov(aX, bY)}{\sqrt{Var(aX)}\sqrt{Var(bY)}} = \frac{ab}{|ab|}Corr(X, Y)$$

- $$Cov(X, X) = E[(X - \mu_X)(X - \mu_X)] = Var(X)$$

X, Y, Z, W

- : 확률변수

$$Cov(X + Y, Z) = E[(X + Y - (\mu_X + \mu_Y))(Z - \mu_Z)]$$

$$= E[(X - \mu_X)(Z - \mu_Z)] + E[(Y - \mu_Y)(Z - \mu_Z)] = Cov(X, Z) + Cov(Y, Z)$$

$$Cov(X + Y, Z + W) = Cov(X + Y, Z) + Cov(X + Y, W)$$

$$= Cov(X, Z) + Cov(Y, Z) + Cov(X, W) + Cov(Y, W)$$

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

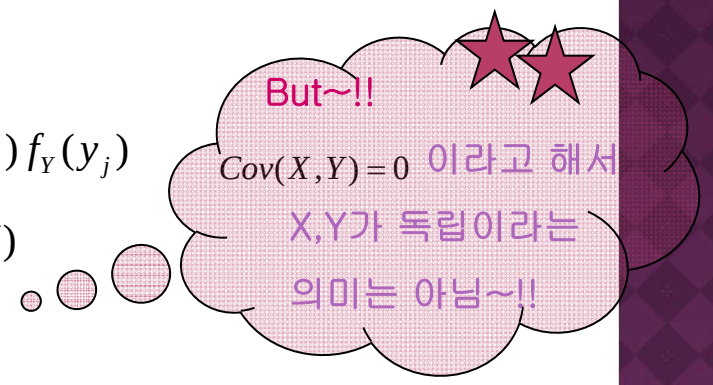
A, B $\quad P(AB) = P(A)P(B)$

$$X : x_1, x_2, \cdots, x_m$$
$$Y : y_1, y_2, \cdots, y_n$$

X, Y

$$f(x_i, y_j) = P[X = x_i, Y = y_j] = P(X = x_i)P(Y = y_j) = f_X(x_i) f_Y(y_j) \quad i = 1, \cdots, m \quad j = 1, \cdots, n$$

$$E(XY) = \sum_j \sum_i x_i y_j f(x_i, y_j) = \sum_j \sum_i x_i y_j f_X(x_i) f_Y(y_j)$$
$$= \sum_j y_j f_Y(y_j) + \sum_i x_i f_X(x_i) = E(X)E(Y)$$
$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0$$



누적확률

- X ~ 확률분포 $\quad f(x) = P[X = x]$
- 누적확률 : $\quad F(c) = P[X \le c] = \sum_{x \le c} f(x)$

x	$f(x)$	$F(x)$
1	0.07	0.07
2	0.12	0.19
3	0.25	0.44
4	0.28	0.72
5	0.18	0.9
6	0.1	1

$$P[X = 3] = F(3) - F(2) = 0.44 - 0.19 = 0.25$$

$$P[X \ge 3] = 1 - F(2) = 1 - 0.19 = 0.81$$

Thank You!