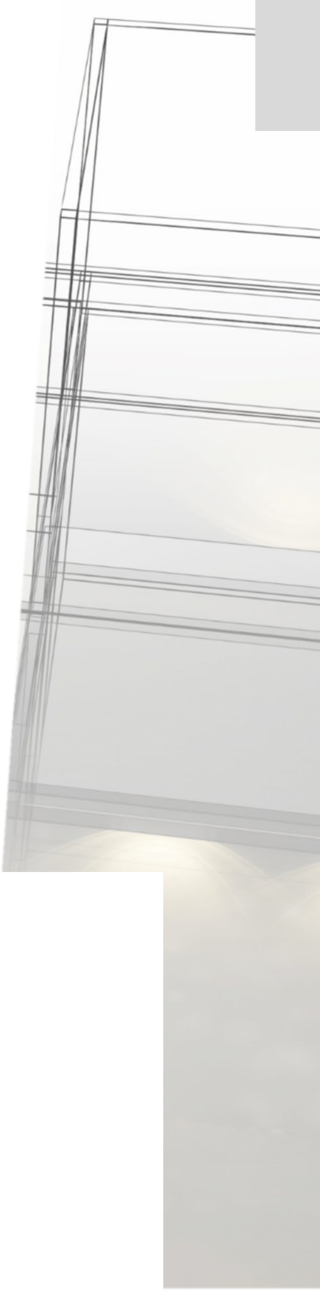
다변량 확률분포

통계·데이터과학과
이긍희 교수

학습목표

1. 결합분포를 이해할 수 있다.
2. 공분산과 상관계수를 이해할 수 있다.
3. 조건부분포를 이해할 수 있다.
4. 다변량 정규분포를 이해할 수 있다.
5. 다항분포를 이해할 수 있다.



01

10강 다변량 확률분포

결합분포와 주변분포

결합분포

◆ 결합분포 : 두 개 이상의 확률변수를 함께 고려한 확률분포 → 다변량 확률분포

결합분포

- $P(X \geq 172) = 0.5,\ P(Y \geq 68) = 0.5$

- 다음 두 경우 두 변수의 관계는?

  $P(X \geq 172, Y \geq 68) = 0$

  $P(X \geq 172, Y \geq 68) = 0.25$

이산형 확률변수의 결합분포

◆ X와 Y의 결합분포 : 결합 확률질량함수

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

이산형 확률변수의 결합분포

◆ X와 Y의 결합분포 : 결합 확률질량함수

X \ Y		Y			
		y_1	y_2	$\cdots$	y_n
X	x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\cdots$	$p(x_1, y_n)$
	x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\cdots$	$p(x_2, y_n)$
	$\vdots$			$\vdots$	
	x_m	$p(x_m, y_1)$	$p(x_m, y_2)$	$\cdots$	$p(x_m, y_n)$

이산형 확률변수의 주변분포

◆ X와 Y의 주변분포

$$P(X = x_i) = \sum_{j=1}^{n} p(x_i, y_j)$$

$$P(Y = y_j) = \sum_{i=1}^{m} p(x_i, y_j)$$

결합분포와 주변분포의 예

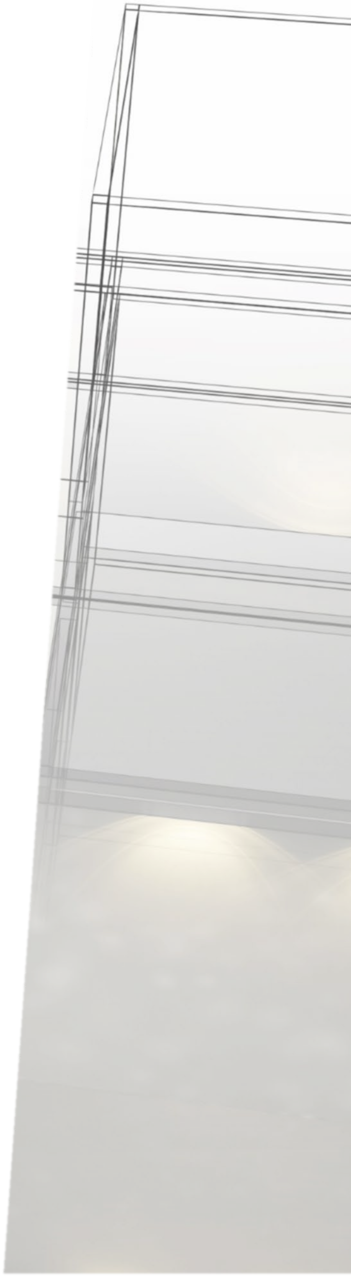
예

500원, 100원, 50원 동전 던지는 실험

X : 500원 앞면 여부, Y : 총 앞면의 수

X, Y의 결합분포와 주변분포는?

HHH	$\longrightarrow (X=1,\ Y=3)$	THH	$\longrightarrow (X=0,\ Y=2)$
HHT	$\longrightarrow (X=1,\ Y=2)$	THT	$\longrightarrow (X=0,\ Y=1)$
HTH	$\longrightarrow (X=1,\ Y=2)$	TTH	$\longrightarrow (X=0,\ Y=1)$
HTT	$\longrightarrow (X=1,\ Y=1)$	TTT	$\longrightarrow (X=0,\ Y=0)$



이산형 확률변수의 주변분포 예

예

500원, 100원, 50원 동전을 던지는 실험
X, Y의 결합분포와 주변분포는?

이산형 확률변수의 주변분포 예

예

500원, 100원, 50원 동전을 던지는 실험
X, Y의 결합분포와 주변분포는?

$HHH \longrightarrow (X=1,\ Y=3)$	$THH \longrightarrow (X=0,\ Y=2)$
$HHT \longrightarrow (X=1,\ Y=2)$	$THT \longrightarrow (X=0,\ Y=1)$
$HTH \longrightarrow (X=1,\ Y=2)$	$TTH \longrightarrow (X=0,\ Y=1)$
$HTT \longrightarrow (X=1,\ Y=1)$	$TTT \longrightarrow (X=0,\ Y=0)$

연속형 확률변수의 결합분포

◆ (X, Y) 의 결합 확률밀도함수 : $f(x, y)$

① 모든 (x, y) 에 대하여 $f(x, y) \geq 0$

② $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

③ $P((X, Y) \in R) = \int\int_R f(x, y) dx dy$

연속형 확률변수의 주변분포

◆ (X, Y)의 결합 확률밀도함수 : $f(x, y)$

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx$$

연속형 확률변수 결합분포의 예

예

(X, Y)의 결합 확률밀도함수

$$f(x, y) = 2x - 4xy + 2y, 0 \le x \le 1, 0 \le y \le 1$$

(1) $P(X \le \frac{1}{2}, Y \le \frac{1}{2})$일 확률은?

연속형 확률변수 결합분포의 예

예

(X, Y) 의 결합 확률밀도함수

$$f(x,y) = 2x - 4xy + 2y, 0 \le x \le 1, 0 \le y \le 1$$

(2) X 의 확률밀도함수 $f_X(x)$ 는?

02

10강 다변량 확률분포

공분산과 상관계수

공분산

◆ 공분산(covariance)

두 변수간 관계 파악

- $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$

공분산

- 이산형 확률변수 :

$$Cov(X,Y) = \sum_{i,j}\left(x_i - E(X)\right)\left(y_j - E(Y)\right)p(x_i, y_i)$$

- 연속형 확률변수 :

$$Cov(X,Y) = \iint_{x,y}\left(x - E(X)\right)\left(y - E(Y)\right)f(x,y)dxdy$$

공분산의 예

예

두 확률변수 X, Y의 공분산을 구하시오.

X \ Y	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

공분산의 특성

- $Cov(X, a) = 0$
- $Cov(aX, bY) = ab\,Cov(X, Y)$
- X, Y 독립 $\rightarrow Cov(X, Y) = 0$
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$

상관계수

- $\rho = Corr(X, Y) = \dfrac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$

상관계수의 특성

- $-1 \leq \rho \leq 1$
- $\rho > 0 \rightarrow X, Y$ 간 양의 선형관계
- $\rho < 0 \rightarrow X, Y$ 간 음의 선형관계
- $\rho = 0 \rightarrow X, Y$ 간 선형관계 없음

상관계수의 예

예

두 확률변수 X, Y의 상관계수를 구하시오.

X \ Y	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

상관계수의 예

예

두 확률변수 X, Y의 상관계수를 구하시오.

03

10강 다변량 확률분포

조건부 분포

조건부 분포

◆ Y값이 주어졌을 때, X의 확률분포 : 이산형

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i\, ,\ Y = y_j)}{P(Y = y_j)}$$

조건부 분포

◆ Y값이 주어졌을 때, X의 확률분포 : 연속형

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

조건부 분포의 예

예

$Y = 2$일 때의 X의 확률분포(조건부 분포)
$P(X = x \mid Y = 2)$를 구하시오.

X \ Y	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

조건부 기댓값

- $E(X|Y=y_j) = \sum_{i=1}^{n} x_i P(X=x_i|Y=y_j)$
- $E(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx$

조건부 기댓값의 예

예

조건부 확률분포의 기댓값은?

X	0	1
$P(X = x \mid Y = 2)$	1/3	2/3

04

10강 다변량 확률분포

다변량 정규분포

이변량 정규분포

- 정규분포를 따르는 2개 확률변수의 결합분포

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

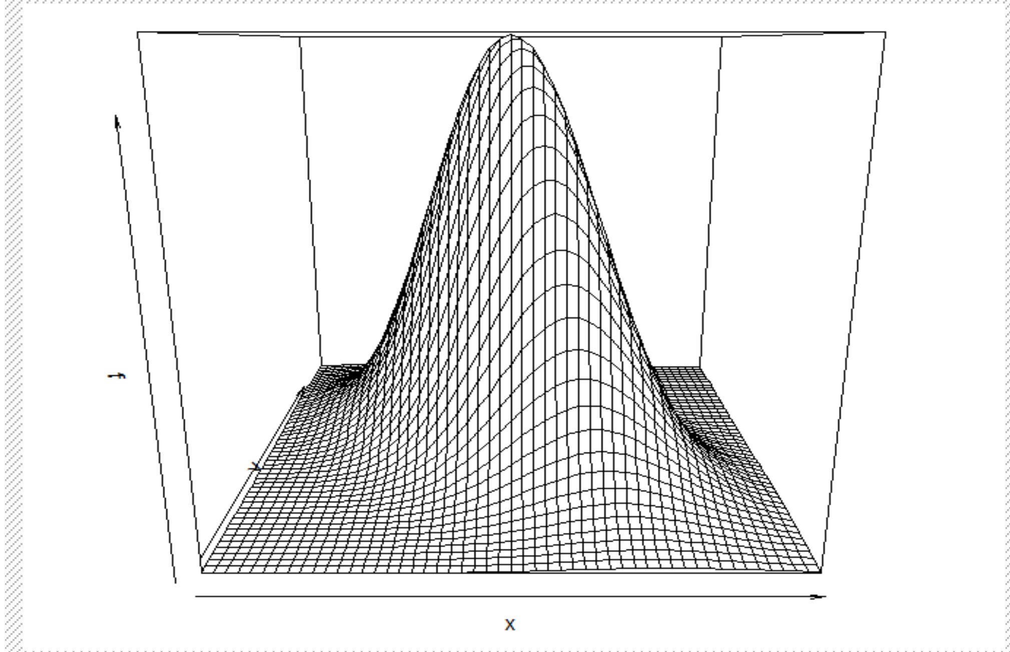
이변량 정규분포

- 확률밀도함수

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

이변량 정규분포

◆ 이변량 정규분포의 확률밀도함수



이변량 정규분포의 예

예

$$(X,Y) \sim N_2\left[\begin{pmatrix}3\\5\end{pmatrix},\begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}\right]$$

(1) X의 확률분포는?

이변량 정규분포의 예

예

$$(X,Y) \sim N_2\left[\begin{pmatrix} 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}\right]$$

(2) X , Y의 상관계수는?

조건부 정규분포

◆ 조건부 정규분포

$$Y|X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right)$$

이변량 정규분포의 예

예

$$(X,Y)\sim N_2\left[\begin{pmatrix}3\\5\end{pmatrix},\begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}\right]$$

$X=0$일 때 Y의 조건부 분포는?

다변량 정규분포

◆ 정규분포를 따르는 p개 확률변수 결합분포 :

$$(X_1, X_2, \cdots, X_p) \sim N_p(\boldsymbol{\mu}, \Sigma)$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

다변량 정규분포

◆ 결합 확률밀도함수

$$f\left(x_1, x_2, \dots, x_p\right) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

05 다항 분포

멀티누이 분포

- 배반적 두가지 결과 → 베르누이 분포
- 배반적 여러 결과 → 멀티누이 분포
- 멀티누이 분포(multinoulli distribution)
  : 일반화된 베르누이 분포, 범주형 분포

멀티누이 분포

- $P(X_i = 1) = p_i \;\; , P(X_i = 0) = 1 - p_i$
- $P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$

  $x_i \; = 0, 1 \quad x_1 + x_2 + \cdots + x_k = 1$

다항분포

- 각 시행에서 나올 결과가 k개인 사건을 독립적으로 n번 시행. 각 범주별 발생 횟수들$(X_1, X_2, \cdots, X_k)$
  → 다항분포(multinomial distribution)

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

다항분포의 예

예

어떤 식물의 유전자 형태 AA, Aa, aa로 구분. 각각의 확률, 0.25, 0.5, 0.25. 10개를 임의로 선택, AA, Aa, aa가 각각 3, 5, 2일 확률은?

학습정리

정리하기

- 두 개 이상의 확률변수의 분포를 결합분포라고 하며, 결합분포 구성 확률변수의 분포를 주변분포라고 한다.
- 두 확률변수간의 선형적 연관성은 공분산, 상관계수로 측정한다.
- 조건부 분포는 한 확률변수의 값이 주어졌을 때 다른 확률변수의 확률분포이다.

학습정리

- 다변량 정규분포는 정규분포를 따르는 2개 이상 변수의 결합 확률분포이다.

- 멀티누이 분포와 다항분포는 각각 베르누이 분포와 이항 분포에서 범주가 여러 개인 형태로 확장한 것이다.

다음시간에는

수고하셨습니다.

10강 다변량 확률분포

11강 표본분포1